# A QUASI-EXPERIMENTAL STUDY OF THE EFFECT OF MATHEMATICS PROFESSIONAL DEVELOPMENT ON STUDENT ACHIEVEMENT

by

**Zahid Kisa**

BS, Mathematics Education, Bogazici University, 2005

MA, Educational Sciences, Bogazici University, 2008

Submitted to the Graduate Faculty of

The School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Learning Science and Policy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Zahid Kisa

It was defended on

July 21, 2014

and approved by

Dr. Mary Kay Stein, Professor, Learning Sciences and Policy

Dr. Lindsay Clare Matsumura, Associate Professor, Learning Sciences and Policy

Dr. Michael P. Marshal, Associate Professor, Psychiatry and Pediatrics

Dissertation Advisor: Dr. Richard Correnti, Associate Professor, Learning Sciences and Policy

**A QUASI-EXPERIMENTAL STUDY OF THE EFFECT OF MATHEMATICS PROFESSIONAL DEVELOPMENT ON STUDENT ACHIEVEMENT**

Zahid Kisa, PhD

University of Pittsburgh, 2014

Over the past couple of decades, teacher effectiveness has become a major focus to improve students' mathematics learning. Teacher professional development (PD), in particular, has been at the center of efforts aimed at improving teaching practice and the mathematics learning of students. However, empirical evidence for the effectiveness of PD for improving student achievement is mixed and there is limited research-based knowledge about the features of effective PD not only in mathematics but also in other subject areas. In this quasi-experimental study, I examined the effect of a Math and Science Partnership (MSP) PD on student achievement trajectories. Results of hierarchical growth models for this study revealed that content-focused (Algebra1 and Geometry), ongoing PD was effective for improving student achievement (relative to a matched comparison group) in Algebra1 (both for high and low performing students) and in Geometry (for low performing students only). There was no effect of PD on students' achievement in Algebra2, which was not the focus of the MSP-PD. By demonstrating an effect of PD on student achievement, this study contributes to our growing knowledge base about features of PD programs that appear to contribute to their effectiveness. Moreover, it provides a case study showing how the research design might contribute in important ways to the ability to detect an effect of PD -if one exists- on student achievement. For example, given the data I had from the district, I was able to examine student growth within all

Algebra 1, Geometry and Algebra 2 courses, while matching classrooms on aggregate student characteristics and school contexts. This allowed me to eliminate the potential confound of curriculum and to utilize longitudinal models to examine PD effects on students' growth (relative to a comparison sample) for matched classrooms. Findings of this study have implications for educational practitioners and policymakers in their efforts to design and support effective PD programs in mathematics, and these features likely transfer to the design of PD in all subject areas. Moreover, for educational researchers this study suggests potential strategies for demonstrating robust research-based evidence for the effectiveness of PD on student learning.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## ACKNOWLEDGMENTS

During the time that I have been a graduate student at the University of Pittsburgh, a number of people – friends, colleagues, professors, and family –have contributed to making this an enriching professional and personal experience for me. I am deeply grateful for their support, and want to express my sincere thanks to as many of them as possible.

First, I want to thank my advisor Richard Correnti, who has been a mentor in every way. He knew exactly what kind of guidance I needed. More importantly, he has been a great role model for me. He is a selfless, reachable and a considerate person all the time. He demonstrated how expertise in both instructional policy and quantitative analyses creates robust educational research studies. His true and generous guide has been very helpful for me as I developed my scholarship and professional skills.

I am grateful for the profound guidance from my committee members: Mary Kay Stein, Lindsay Clara Matsumura, and Michael Marshal. Learning from and having discussion with a scholar like Mary Kay, whose research has made substantial contributions to the field is an invaluable experience. I am grateful to Lindsay for her elegant critiques and constructive feedback, which helped to strengthen my dissertation. Michael's careful look into the statistical analyses in my dissertation resulted in important refinements for the presentation of the findings.

Many faculty at the School of Education have been supportive and inspiring during my career at Pitt. I especially want to thank Phillip Herman, Jennifer Russell, Kevin Kim, and Feifei

Ye. I also want to thank my LSAP colleagues and friends especially my office mate Elaine Wang.

I want to thank my parents, my sister, nephew and nieces back in Turkey. Thank you for your understanding when I had to miss - for the sake of achieving my goal- celebrations, holidays, birthdays, and special moments that I could have shared with you. I also want to thank my parents-in-law for their support. I am lucky since I have the most welcoming parents-in-law a person could ask for.

Unquestionably, my greatest thank you is to my wife, Miray Tekkumru Kisa, whose contribution to my career is invaluable. We have been faced with many challenges, have overcome many obstacles and have reached our goals by always being together! I look forward to all of our future endeavors and adventures together. Thank you for being a part of my life!

# 1.0 INTRODUCTION

Over the past couple of decades, many efforts have been made to reform mathematics education in the United States. However, results from national and international assessments indicate that reforms haven't been successful in improving U.S. students' mathematics achievement (Ball, Hill, & Bass, 2005). Scores on the 2013 National Assessment of Educational Progress (NAEP), indicate that overall only 26% of the nation's twelfth-graders were at or above the proficient level (National Center for Education Statistics [NCES], 2013). On an international assessment, PISA (The Program for International Student Assessment) administrated in 2009, U.S. students scored lower than the OECD average. This continued the trend of U.S. students scoring lower than the OECD average in 2003 and 2006 administrations of PISA (Epstein, & Miller, 2011).

Concerns about the performance of U.S. students on such mathematics assessments have resulted in policymakers paying increased attention to issues of teacher quality and teacher effectiveness (Hill, Rowan, & Ball, 2005). At the same time, realization of the fact that the success of standards-based educational reform in mathematics relies on the effectiveness of teachers, has also played a role resulting in increased attention to the issue of teacher effectiveness (Darling-Hammond, 1999). Furthermore, one of the important prerequisites for teachers to carry out the requirements of standards-based educational reform (e.g. opening their classroom to wider mathematical participation, helping students to appreciate mathematical reasoning, and to understand the meaning of mathematical ideas and procedures), is deep

mathematical content knowledge as well as deep pedagogical content knowledge (Hill & Ball, 2004).

However, multiple studies have revealed that many teachers are not ready to implement teaching practices based on ambitious education reform and teachers lack essential content knowledge for teaching mathematics (Ball et al., 2005). As a result, teacher professional development (PD), in particular, has been at the center of efforts aimed at improving teaching effectiveness (Garet, Porter, Desimone, Birman, & Yoon, 2001; Sykes & Darling-Hammond, 1999). All the while, changes in student achievement remains the primary criteria for demonstrating teacher effectiveness, and therefore demonstrating the value of PD programs aimed at improving teaching effectiveness. States and school districts are currently providing PD programs on a wide scale, several with federal funding support, in order to improve the quality of their teachers (Hill et al., 2005).

One such effort is the Math-Science Partnership (MSP) program created by the National Science Foundation (NSF) with the goal of improving student achievement in mathematics and science through providing content-focused PD. The MSP grant program encourages institutions of higher education, and K-12 schools to work together in order to increase the quality of mathematics and science instruction and student learning in mathematics and science fields by providing high quality PD for teachers. NSF has spent approximately $800 million to fund various MSP projects across the US since 2002. The U.S. Department of Education has also supported efforts to provide PD for teachers in the STEM areas and has spent over $70 million for this purpose since 2003. MSP programs are being implemented in at least 39 states in the United States (National Science Foundation [NSF], 2010).

## 1.1 FOCUS OF THE STUDY

In this longitudinal quasi-experimental study, I examined the effect of one district's PD program on changes in student achievement over a year. This PD program was designed and implemented with support from the Math-Science Partnership (MSP) program[1]. The Math and Science Partnership professional development program examined in this study (hereafter referred to as MSP-PD) was designed and implemented as a product of a collaborative effort of a non-profit organization, a university and an urban school district located in the Northeastern U.S. with about $760,000 MSP grant support. The MSP-PD included a 2-week summer institute and 6 follow-up sessions during the school year. On average MSP-PD teachers were provided 110 hours of PD (80 hours summer institutes, 30 hours follow-up sessions).

The pedagogy of the MSP-PD was inquiry-based mathematics (e.g. hands on learning experiences, active participation of teachers) The MSP-PD focused on teachers' actively engaging with each other and with course facilitators in doing mathematics in the topic areas algebra1 and geometry. It modeled the teaching that district leaders were hoping their mathematics teachers would adopt with their students. Teachers were both working on content as mathematicians and engaging in mathematics as a discipline. They reflected on their own disciplinary experiences as learners and discussed implications of creating similar learning opportunities for their students. Moreover, the content of the MSP-PD was closely linked with the curriculum.

[1] PD programs designed and implemented with support from the MSP program are similar only in terms of being content focused and having a partnership model. Other features of the PD programs can be different. Thus, MSP-PD examined in this study may not be representative of the MSP programs in general.

### 1.1.1 Research Questions

The theory of action of the MSP-PD was as follows: By engaging in the doing of mathematics, teachers would improve or reinvigorate their content knowledge and be motivated (and better prepared) to facilitate environments where their students engaged in doing mathematics. Based on the theory of action, we would expect some improvements in teacher and student outcomes as a result of the MSP-PD. Thus, this study sought to examine both proximal and distal outcomes as evidence for the efficacy of the MSP-PD. First, I examined whether the 2-week summer institute resulted in changes in teachers' algebra content knowledge.

**RQ1-a** For teachers who attended the MSP-PD, was their mean post-test score on the knowledge of algebra teaching assessment (KAT) significantly higher than their mean pre-test score?

While this question helps us understand whether treated teachers improved, it is also important to know how different treated teachers were from a group of comparison teachers at pre-test. Thus, I also compared the scores of the treated teachers to a group of comparison teachers in the district.

**RQ1-b** Was there a significant difference between MSP-PD teachers' mean KAT pre-test score and comparison group teachers' mean KAT score?

Based on its theory of action, the effect of the MSP-PD on this proximal teacher outcome would have more meaning if it in turn leads to improvements in student learning. Thus, I also examined whether MSP-PD had an effect on growth in student learning relative to a matched group of comparison students.

**RQ 2** To what extent did the MSP-PD influence students' trajectories of mathematics achievement relative to a matched comparison group for each math course over one school year?

Finally, I hypothesized that the effect of the MSP-PD on students' trajectories and final status on the curriculum-based mathematics assessment varied by the course they taught and the degree to which it was aligned with the PD. Thus, I expected that the MSP-PD produced a higher growth rate for students in algebra1 and geometry courses because algebra1 and geometry topics were more aligned with the focus of the PD.

**RQ 3** Did the size of the MSP-PD effect vary for different courses in predictable ways?

## 1.2 SIGNIFICANCE OF THE STUDY

Teachers are crucial to students' opportunities to learn mathematics (Ball et al., 2008). They determine how much time will be devoted to a subject, set and communicate standards and expectations, and decide which topics will be the focus of student learning (Hawley & Valli, 1999; Schwille et al., 1983). Studies examining student achievement in mathematics have found that substantial differences in achievement between students are attributable to differences among teachers. For example, children assigned to three effective teachers in a row scored 50 percentile points higher than children who were assigned to three ineffective teachers in a row (Sanders & Rivers, 1996). Other studies, meanwhile, demonstrate that the cumulative effects of being taught by consecutive highly effective teachers can substantially eliminate differences in student achievement that are due to family background (Rivkin, Hanushek, & Kain, 2005).

The fact that teachers play such a crucial role for students' learning highlights the importance of PD as a vital tool for reform because of the potential to improve students' achievement by increasing teaching effectiveness. The MSP-PD examined in this study underscores this point. At the time of this study, this MSP-PD was one of the primary efforts by

the district to improve the mathematics achievement of its high school students. This study, thus, contributes valuable knowledge about the effectiveness of this key tool, MSP-PD, for improving student mathematics achievement.

Examining the effectiveness of one particular PD program is an important goal for this study and a potential contribution to growing research knowledge in mathematics, but this study also contributes to the current research-based knowledge about the effect of PD on student achievement across all subject areas. Considering that amount of money spent annually on PD, including MSP grants (Birman & Porter, 2002; Miles, Odden, Archibald, & Fermanich, 2002), and its widespread use as a tool of reform for improving students learning (Cohen & Hill, 2000; Hawley & Valli, 1999; Knapp, 2003), producing research-based knowledge about the effectiveness of PD in any subject could lead to generalizations for the field about effective PD designs. Additionally, recent calls from policymakers for evidence-based education research further elevate the importance of understanding the effects of PD on student learning (Birman & Porter, 2002; Coalition for Evidence-Based Policy, 2003), and also how education research is optimally designed to examine effects of PD programs.

## 1.3 CONTRIBUTION FOR THE KNOWLEDGE-BASE ABOUT THE EFFECTIVENESS OF PD

It is generally accepted that effective professional development (PD) can improve teachers' knowledge and thus create change in their instructional practices (Arbaugh & Brown, 2005). This makes it an important tool for improving students learning (Borko, 2004; Correnti, 2007; Desimone, 2009; Sykes & Darling-Hammond, 1999). However, relative to PD literature there is

very small number of experimental/quasi-experimental studies investigating effect of PD on students' achievement. Ball et al. (2008) searched for peer-reviewed research and national reports that would offer high-quality evidence regarding the impact of PD programs on students' mathematics achievement. They identified only eight high-quality empirical studies that examined effects of PD on students' mathematics achievement. Among these studies only one study (Chapin, 1994) provided clear evidence for the effectiveness of mathematics PD. Two studies (Carpenter, Fennema, Peterson, Chiang, & Loaf, 1989; Saxe, Gearhart, & Nasir, 2001) found mixed effect and rest of the studies found no effect of mathematics PD for improving student achievement (See Ball et al., 2008). Moreover, from three experimental/quasi-experimental studies, which were conducting in recent years, Newman et al. (2012) found very small effect of mathematics PD (ES=0.05). McMeeking and her colleagues (2012) found an effect of mathematics PD when teachers attended the PD several times. On the other hand, Garet et al. (2010; 2011), found no effect of a mathematics PD on students' learning after the first and second years of implementation. Thus, in the PD literature there is currently weak empirical evidence for the effectiveness of mathematics PD for improving students' achievement.

The link between PD and student achievement is weak in other subject areas as well. In the broader PD literature, what we know about the particular characteristics of effective PD is mostly based on theoretical and practical advice, expert understanding of what works, correlational survey studies, and case studies (American Educational Research Association [AERA], 2005; Scher & O'Reilly, 2009). PD that is content focused and intense, close to teachers' classroom practice, aligned with standards for teaching, lasting for a longer duration, requiring active participation of teachers, providing teachers hands on learning opportunities, modeling examples of intended practices, providing ongoing support like coaching and

mentoring are characteristics that have been identified as some of the features of effective PD (Ball & Cohen, 1999; Cohen & Hill, 2000; Supovitz, Mayer, & Kahle, 2000; Hawley & Valli, 1999; Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009).

While a long list of effective PD features was developed mostly from theory, experimental/quasi-experimental studies, however, have provided empirical evidence only for two PD features. These studies provided evidence for the effectiveness of content focused PD and mixed evidence for the effectiveness of number of contact hours/duration of the PD (Blank & de las Alas, 2009; Scher & O'Reilly, 2009; Yoon, Duncan, Scarloss, & Shapley, 2007). Thus, despite a long history of educational researchers theorizing about specific feature of PD that promote student learning, empirical evidence supporting those assertions remains very limited and centers on providing PD with a content focus (Fishman, Marx, Best, & Tal, 2003). Thus, there is a need for rigorous studies and evaluations of PD programs in order to learn the extent to which PD influences student achievement. By examining different features of specific PD programs, case studies can provide empirical evidence for the effectiveness of PD for improving student achievement. More importantly, accumulation of these studies can help building a knowledge base about what forms of, or approaches to PD lead to improvement in student learning.

For example, this study uses a rigorous quasi-experimental design to provide evidence for the effectiveness of a content-focused, ongoing PD on students' mathematics achievement. Although it is a mathematics PD program, the main features of the MSP-PD could generalize to help build our knowledge base. Hence, findings of this study not only inform those interested in this specific MSP-PD, but the broader PD literature about the effectiveness of PD for influencing

students' achievement and the literature about which specific features of PD programs seem to be effective for improving student achievement.

## 1.4 ISSUES IN PD RESEARCH AND THEIR IMPLICATIONS FOR THIS STUDY

One of the main reasons for the lack of strong evidence for the effectiveness of PD in the literature is simply the shortage of experimental/quasi-experimental studies examining the effect of the PD on student achievement (Ball et al., 2008). However, the shortage of studies doesn't explain *why* PD studies have produced mixed evidence for the effectiveness of PD. Experimental/quasi-experimental studies have found some positive effects, some mixed effects, some null effects (i.e. no effect) and even some negative effects of PD on student achievement (Ball et al., 2008; Blank & de las Alas, 2009; Clewell, Campbell, & Perlman, 2004; Kennedy, 1998; Scher & O'Reilly, 2009; Yoon et al., 2007). In general, the mixed evidence for the effectiveness of PD raises questions about whether PD is the most effective tool to improve student achievement (Harris & Sass, 2011).

In order to reach a conclusion about the effectiveness of PD, however, one should be confident that the observed research effects were, in fact, due only to the PD itself. In PD research, there are many obstacles, leading to multiple alternative explanations, that could influence our findings about the effects of PD on student learning (Scher & O'Reilly, 2009). For example, the curriculum, implementation of the PD, teacher mobility and student mobility could all confound observed effects of the PD. Moreover, research-related factors, such as the nature of student data (cross sectional vs. longitudinal), and alignment between PD and assessment measures, might also have a role in an observed null effect of PD on achievement (Blank & de

las Alas, 2009; Kennedy, 1998; Yoon et al., 2007). These factors might confound or change our inferences about the effect of PD on student achievement. Most PD studies have examined the effect of PD on student achievement without explicitly attending to the role these factors can have on the observed effect of PD. In most cases, research studies have also had limited control over the influence of such factors due to the complexity of educational settings. The challenges researchers face in collecting research-based evidence clouds our ability to make strong inferences and to be certain about the effects of PD on teacher knowledge, instruction and student learning (Ball et al., 2008).

For PD designers, prior PD studies can be informative about which features of PD tend to have a greatest influence of student outcomes. Similarly, educational researchers can benefit from observing patterns across prior PD studies to have an idea about which design features of the research were helpful for making strong inferences about the effect of a PD on student achievement. By examining patterns across recent experimental/quasi-experimental studies and linking these findings with the knowledge base about quantitative research methodology, this study compiled a list of features that will aid in making strong inferences about the effectiveness of PD for improving student achievement. These design elements are 1) implementing carefully designed experimental/quasi-experimental studies; 2) attending to the fidelity of implementation of PD; 3) examining intermediate teacher outcomes of PD such as teacher knowledge or practice; 4) using proximal and student learning measures which are sensitive to and aligned with the PD; 5) using longitudinal data that allow researchers to employ growth modeling; and 6) trying hard to make populations of students taught by treated teachers as similar to students of comparison teachers as possible. We can use this list to understand the features available in particular

contexts, like our study of one MSP-PD, but also to help educational researchers plan for future studies.

Addressing all of these elements of research design in a single study is challenging and the degree to which researchers can incorporate these design features involves trade-offs. However, by attending to these principles it is possible to make stronger inferences that observed effects on student achievement are due to the PD itself and not from other confounding variables. Perhaps more importantly, the design of research studies is important for interpreting null findings. In poorly designed research studies it is difficult to know whether null findings were due to the research design or the PD itself. Well-designed and adequately powered research studies will help advance generalizations researchers can make when examining patterns across studies because null findings are possible to interpret. In this study, I employed some of these design elements listed above While it is not possible to test the effectiveness of these design features in a single study, this study serves as a case study showing how the research design might contribute in important ways to the ability to detect an effect of PD on student achievement. This study provides an opportunity to reflect on potential features of research designs examining effects of PD on student achievement and thus contribute to a discussion about methods for designing studies of PD.

# 2.0 LITERATURE REVIEW

In order to review the existing knowledge base for the effectiveness of PD, I examined experimental/quasi-experimental studies that used student learning or achievement as an outcome to assess the effectiveness of PD. Relative to the number of PD studies in the literature, since there are very few experimental and quasi-experimental studies specifically examining the link between mathematics PD and students' achievement (Ball et al., 2008; Garet et al., 2010; 2011; McMeeking et al,, 2012; Newman et al., 2012), I explored the broader PD literature in order to summarize the current research-based evidence for the effectiveness of PD for improving student achievement. Given the relative scarcity of research-based evidence for the effectiveness of PD on student achievement, any knowledge produced by experimental/quasi-experimental case studies would be a contribution to the field.

There are several comprehensive review studies that have examined high-quality experimental/quasi-experimental studies. These review studies began by examining vast numbers of studies but whittled their review down to just a handful of high-quality experimental and quasi-experimental studies that examined the effect of PD on student achievement (See Table 1). For example, Yoon et al. (2007) analyzed over 1,300 studies and evaluation reports, and identified only 9 high-quality experimental or quasi-experimental studies that evaluated PD impacts on student achievement. These review studies have synthesized evidence for the overall effectiveness of PD on student achievement across this small set of high-quality PD studies.

They also compared different features of PD programs in order to infer the effectiveness of specific features. Findings from these review studies have helped to catalog the evidence base for the effectiveness of PD produced by PD studies with rigorous designs.

**Table 1.** Number of high quality experimental/quasi-experimental PD studies identified by the review studies

| Review Study | Number of scanned PDs | Number of high quality PDs |
|---|---|---|
| Kennedy (1998) | 93 | 10 |
| Clewell et al. (2004) | 400 | 18 |
| Yoon et al. (2007) | 1300 | 9 |
| Ball et al. (2008) | - | 8 |
| Blank & de las Alas (2009) | - | 16 |
| Scher & O'Reilly (2009) | 145 | 8 |

In addition to these review studies, I also investigated experimental/quasi-experimental studies examined the effect of PD on student achievement which were published after 2007 since the reviews had yet to include these more recent studies in their analyses. In this section, I shared empirical evidence provided by these studies for the overall effectiveness of PD and for the specific features of PD for improving student achievement.

## 2.1 THE EMPIRICAL EVIDENCE FOR THE EFFECTIVENESS OF PD ON STUDENT ACHIEVEMENT IS MIXED

### 2.1.1 Evidence for the overall effectiveness of PD

The review studies that have examined the effects of experimental/quasi-experimental PD studies on student achievement have found mixed evidence for the effectiveness of PD programs. Across the studies the overall PD effect on student achievement was small to moderate [2](Ball et al., 2008; Blank & de las Alas, 2009; Clewell et al., 2004; Kennedy, 1998; Scher & O'Reilly, 2009; Yoon et al., 2007). For example, Scher and O'Reilly (2009) indicated that students whose teachers participated in PD achieved, on average, 0.14 standard deviations higher on mathematics achievement tests and 0.13 standard deviations higher on science assessments compared to students whose teachers did not participate in the PD programs. Moreover, there is considerable variation between studies in terms of the effect sizes of the PD programs (See Table 2). For example, 25 percent of the effect sizes in Blank & de las Alas (2009) study were negative, 56 percent of the effect sizes were small, and only 2 of 104 total effect sizes were large.

[2] According to the Cohen's d standard guidelines, d=0.2 is considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size.

**Table 2.** Mean effect sizes and ranges for PD effects on student achievement calculated in review studies

| Review Study | Subject Matter | Mean ES | Range of ES |
|---|---|---|---|
| Kennedy (1998) | Math & Science | 0.29 (Small-Moderate) | (-0.14-0.71) |
| Yoon et al. (2007) | Math, Science, Literature | 0.54 (Moderate) | (-0.53-2.39) |
| Ball et al. (2008) | Math | 0.42 (Moderate) | (-1.34-2.39) |
| Blank & de las Alas (2009) | Math & Science | 0.31 (Small-Moderate) | (-0.19-1.63) |
| Scher & O'Reilly (2009) | Math & Science | 0.14 (Small) | (-0.40-1.40) |

Similarly, recent experimental/quasi-experimental PD studies have also found a wide range of results for the effectiveness of PD on student achievement. The majority of them provided evidence for the overall effectiveness of PD on student achievement (e.g., Biancarosa, Bryk, & Dexter, 2010; Harris et al., 2012; Heller, Daehler, Wong, Shinohara, & Miratrix, 2012; Matsumura, Garnier, & Spybrook, 2013; McMeeking et al., 2012; Powell, Diamond, Burchinal, & Koehler, 2010; Roth et al., 2011; Sailors & Price, 2010). Meanwhile, some studies have found positive effects of the PD on one test or subject, but no effect of the same PD on a different test, or subject (Borman, Boydston, Lee, Lanehart, & Cotner, 2009; Matsumura, Garnier, Correnti, Junker, & Bickel, 2010; Newman et al., 2012). These studies provide mixed evidence because the PD is shown to be effective in just one subject. In contrast, some studies have found that a PD program didn't effect students' achievement at all (Garet et al., 2008; Garet et al., 2010; 2011; Heller, 2012). There is also a study that found negative effects of PD on student achievement (Borman, Gamoran, & Bowdon, 2008).

Although most studies have found positive (but small or moderate) effects of PD on student achievement there are also studies that have found mixed, null, or even negative effects of PD. Overall, the experimental/quasi-experimental studies I reviewed, much like the review studies before them, provide mixed empirical evidence for the effectiveness of PD on student achievement (See Table 3).

### 2.1.2 Evidence for specific features of PD programs.

Finding an overall effect of PD on student achievement is neither as interesting nor as helpful as finding what features of the PD led to improvements in student learning (Scher & O'Reilly, 2009). In reviewing the empirical evidence to date, however, there is even weaker evidence for the effectiveness of specific features of PD programs (Ball et al., 2008; Blank & de las Alas, 2009; Yoon et al., 2007). There are only two specific PD features about which we have some research-based evidence for their effectiveness; content focused PD, and the number of contact hours /duration of the PD.

One of the reasons that there is limited evidence for the effectiveness of PD features is the shortage of experimental and quasi-experimental studies on PD. Because of the small number of experimental/quasi-experimental studies, finding variation between PD programs in terms of features that they adopted is difficult (Ball et al., 2008; Yoon et al., 2007). In most PD studies the effect of one particular PD program was examined. These PD programs, in general, incorporated several design features. For example, in a study conducted by Biancarosa et al. (2010), the PD program was an intense, three-year training with an ongoing coaching component. While review studies have attempted to explore whether there are any systematic differences between specific features of PD programs and their effects on student achievement, due to lack of variation in

features of PD programs, they have only been able to examine the effectiveness of content focused PD and the contact hour/duration of the PD.

Lack of variation between PD programs in terms of the design features that they incorporated was also observed in recent experimental/quasi experimental studies. All of the PD programs were content focused PD and the majority of them incorporated ongoing support via coaching or follow up sessions (See Table 3). As a result, I could only explore whether longer duration, higher contact hour PD programs produced systematically higher student achievement. I could not explore evidence for providing ongoing support because in most cases it was not possible to isolate the effects of coaching or follow up sessions from the effect of longer duration or contact hours.

Considering these two specific features for which the field has accumulated research-based evidence for their effectiveness, there is only convincing empirical evidence for the effectiveness of content-focused PD and the evidence for the effectiveness of the duration/number of contact hours of the PD on student achievement is mixed.

#### 2.1.2.1 Evidence base for the content focused PD

Early review studies achieved the realization that the content of the PD program was important for the effectiveness of the PD on student learning. For example, Kennedy (1998) found that among studies that she examined, PD programs which focused on subject matter knowledge and on students' learning of a particular subject showed the largest effect sizes. Similar to Kennedy's finding, results of the review study conducted by Clewell et al. (2004) concluded that the focus of the PD was critical to observe positive effects on student achievement. Clewell et al. (2004) stated that PD programs that were tied to curriculum, to knowledge of subject matter, and/or to how students learn the subject were more effective in terms of improving student achievement

than PD programs that focused solely on teaching behaviors. They also found that PD programs for teachers of standards/inquiry-based science curricula were associated with higher levels of student achievement. Similarly, Scher and O'Reilly (2009) found that PD programs focused on pedagogical content knowledge were effective in terms of improving students' science and math achievement.

#### 2.1.2.2 Evidence base for duration/ number of contact hours of the PD

For the effectiveness of duration (providing PD over longer time period) and number of contact hours of PD, review studies have found mixed evidence. Kennedy (1998) found no evidence for the effect of higher total contact hours of PD programs on student achievement. However, Yoon et al. (2007) found that PD programs that offered more substantial contact hours (ranging from 30 to 100 hours in total) spread over 6 to 12 months were effective in terms of improving student achievement. Kennedy (1998) found longer duration was effective for mathematics PD but not for science PD, while Scher and O'Reilly (2009) found vice versa.

All of the recent experimental/quasi-experimental studies that have found positive effects of PD on student achievement provided higher total contact hours and were engaged with participants over a longer time period (See Table 3). One PD program which was relatively short (24 hour-long summer PD over five-days) provided no evidence for the effectiveness of the PD (Heller, 2012). However, there were also PD programs that provided more contact hours and were maintained over a longer time period that also found no effects of PD on achievement (Garet et al., 2008; Garet et al., 2010; 2011).

**Table 3.** Duration/number of contact hours of PDs and their observed effects in recent experimental/quasi-experimental studies

| Study | Duration/number of contact hours | Sig. | Component |
|---|---|---|---|
| Matsumura et al., (2013) | Weekly grade level meetings, monthly individually meetings. (2 years) | (+) | Coaching |
| Harris et al., (2012) | 2-day training, ongoing follow up sessions | (+) | Training + follow-up |
| Heller et al., (2012) | 3-hour training, every other week. (14 weeks) | (+) | Training |
| McMeeking et al., (2012) | 2- to 3-week summer training, follow-up sessions across the school year. (Longer than a year) | (+) | Training + follow-up |
| Biancarosa et al., (2010) | 40 hours training in the first year, 10-12 hours training in the subsequent two years and coaching. (3 years) | (+) | Training + coaching |
| Roth, et al., (2011) | 3-week summer training, follow-up sessions across the school year. (1 year) | (+) | Training + follow-up |
| Powell et al., (2010) | 2 day training, coaching. (15 weeks) | (+) | Training + follow-up |
| Sailors & Price (2010) | 2 day training, coaching. (1 year) | (+) | Training + follow-up |
| Garet et al., (2010, 2011) | Summer training, follow up trainings and coaching. (2 years) | (+ . ) | Training + follow-up + coaching |
| Newman et al., (2012) | 10-day training, follow-up training, coaching. (1 year) | (+ . ) | Training + follow-up + coaching |

| | | | |
|---|---|---|---|
| Matsumura et al., (2010) | Weekly grade level meetings, monthly individually meetings. (1 year) | (+ . ) | Coaching |
| Borman, et al., (2009) | 2-day training, annual 1 day training, coaching. (3 years) | (+ . ) | Training + follow-up + coaching |
| Heller (2012) | 24 hours over (5 days) | ( . ) | Training |
| Garet et al., (2008) | Training (during summer and much of the school year), coaching. (1 year) | ( . ) | Training + coaching |
| Borman et al., (2008) | Summer training, coaching | ( _ ) | Training + coaching |
| ( + ) Positive effect; (+ . ) Mixed effect, significant in one outcome but not significant in another outcome; ( . ) Null effect; ( _ ) Negative effect. | | | |

Moreover, Sailors and Price (2010) conducted a systematic comparison of the effectiveness of PD programs of different duration/number of contact hours in a single study. They compared a 2-day workshop with the same workshop followed by classroom-based coaching over the year. Although separating the effect of higher contact hours, longer duration, and coaching was not possible because of the research design, there was evidence for the effectiveness of longer duration PD with higher contact hours (i.e., coaching). In another systematic comparison, Garet et al., (2008) compared the effectiveness of 1) content-focused teacher institute series from summer through much of the school year, 2) the same institute series with coaching component and 3) a control group (no treatment). In this study, while two PD groups-seminar only and seminar with coaching-were similar in terms of having longer duration; the seminar with coaching group was provided higher contact hours than the seminar only group.

Thus, the design of the study made it possible to separate the effect of coaching and longer duration but separating the effect of coaching and high contact hours was not possible. Nevertheless, since they found no effect of both PD programs on student achievement, longer duration PD and longer duration PD with higher contact hours in form of coaching both led to null findings in this study. In all, experimental/quasi experimental studies have provided mixed evidence for the effectiveness of duration/number of contact hours of the PD on student achievement.

The proportion of rigorous research-based studies for the effectiveness of PD is surprisingly very low relative to the large number of PD studies conducted overall. This small set of experimental/quasi-experimental studies has provided empirical evidence only for the effectiveness of content focused PD. There is mixed evidence for the overall effectiveness of PD and for the effectiveness of duration/number of contact hours of the PD.

Considering that in general experimental/quasi-experimental studies produce higher effect sizes than correlational survey studies (Seidel & Shavelson, 2007), and that a publication bias exists toward finding positive effects of PD on student outcomes (Lipsey & Wilson, 1993), it is surprising that experimental/quasi-experimental studies have not shown more robust effects of PD on student achievement. Thus, it is important to understand why experimental/quasi-experimental studies have not provided stronger evidence for the overall effectiveness of PD and for specific features of PD for improving student achievement.

Weak empirical evidence produced by experimental/quasi-experimental studies could be due to the PD itself, which would be an indication that PD is not an effective tool for improving student learning; or it could be due to how we examine the effectiveness of PD on student achievement, which would be an indication that challenges faced by the design of research

studies on PD get in the way of our ability to assess the effects of PD on student learning; or it could be due to a combination of both factors. Below I conducted a thought experiment to understand how much confidence we have in attributing the generally weak evidence of the effects of PD on student learning to the PD programs themselves and how much features of the research design have a role in the lack of strong research-based evidence produced by these studies.

## 2.2 HOW WE SHOULD INTERPRET WEAK EMPIRICAL EVIDENCE FOR EFFECTIVENESS OF PD

The shortage of experimental/quasi-experimental studies plays a role in the weak empirical evidence for the effectiveness of PD. However, it does not explain why PD programs have not produced strong evidence. In other words, having limited empirical evidence for most of the PD features identified in the literature is due to the shortage of PD studies, but having mixed evidence for the effectiveness of duration/number of contact hours of PD or for the overall effectiveness of PD has nothing to do with the shortage of studies. The mixed evidence could raise doubts about whether PD should be a primary tool for improving teaching and learning and whether widely accepted features of effective PD in fact have any influence on student achievement (Harris & Sass, 2011).

In order to reach a conclusion about the effectiveness of PD, one should be confident that the observed research effects were, in fact, due only to the PD itself. In PD research, it is possible to have limited control over factors that potentially confound observed effects of the PD (Scher & O'Reilly, 2009; Yoon et al., 2007). This could lead to multiple alternative explanations for the

effect of PD. I argue that PD studies that found mixed effects of PD within a single study provides us a case showing that the research design and other factors can influence the observed effect of PD. For example, studies that found positive effects of PD on one test or subject, but found no effect of the same PD on a different test, or subject (Borman et al., 2009; Matsumura et al., 2010; Newman et al., 2012) provide an interesting case for understanding how the research design can influence the inferences we make. In a study conducted by Newman and colleagues (2012) the PD had a positive effect on student achievement in mathematics, but there was no effect on students' science achievement. The PD features- focusing on hands-on, inquiry-based instruction, 10-day summer institute with follow-up training and coaching-were similar across the design of the science and mathematics PD programs. Thus, factors that might have a role in observing a mixed effect were not about the PD and PD features. The outcome measure used, for example, might be one factor influencing the observed results because the mathematics test was more specific (SAT 10 problem solving test) while the Science test was general (SAT 10 science test). In addition to PD, teachers were provided with materials as well. Perhaps the math materials were more aligned with the content of the PD than the science materials were. In this study, if the effectiveness of PD were examined only in science, it would be concluded that the PD program was not effective in improving student achievement although the observed effect would not be due only to the design and implementation of the PD program itself.

Other factors could also have a role in the observed effect of PD such as low fidelity of PD implementation, teacher and student mobility, low alignment between study measures and the PD itself, and the effectiveness of the curriculum being used, etc. Most of the PD studies (including high quality experimental/quasi-experimental studies) that examined the effect of PD on student achievement had limited control over the influence of these factors on the observed

effect of PD due to the complex nature of educational settings. As a result, most of the time we have limited confidence in whether the mixed evidence for the effectiveness of PD on student learning is in fact only due to the PD and not to other complications of the research design (Coalition for Evidence-Based Policy, 2003). Thus, the lack of incontrovertible research-based evidence doesn't imply that PD is not an effective tool for improving student achievement. Instead, it implies a need for more experimental/quasi-experimental studies, designed with these complicating factors in mind to produce a more solid knowledge base for the effect of PD on student achievement.

## 2.3 RECOMMENDATIONS FOR DESIGNING EFFECTIVE PD RESEARCH

Given the shortage of PD studies and lack of strong evidence base for the effectiveness of PD, there is a need for rigorous research-based evidence in order to build our knowledge base about the effectiveness of PD. Such PD studies need to make stronger inferences about the role PD plays in producing improved student outcomes and help us better identify the effective features of PD. According to John Stuart Mill, there are three conditions that must be met for making a causal inference, (1) the cause precedes the effect, (2) the cause is related to the effect; if the levels of the cause differ in some systematic way, then there must be corresponding variation in the effect, and (3) we can find no plausible alternative explanation for the effect other than the cause (as cited in Shadish, Cook, & Campbell, 2002; Murnane & Willett, 2010). In complex educational settings, there are several factors, which can influence the observed effect of PD on student achievement. Meeting the third condition for making causal inferences is thus very hard in PD research (Coalition for Evidence-Based Policy, 2003; Cook, 2002; Raudenbush, 2008). In

this section, I provided some suggestions for conducting well-designed PD studies, increasing the odds researchers can find effect of PD when there is an effect of PD (i.e. prevent making type II error) and be more confident in attributing the observed effect to the PD itself. I also reviewed to what extent and how recent high quality experimental/quasi-experimental studies control the factors influencing the observed effect of PD on student achievement through their research design.

### 2.3.1 We should design studies examining the effectiveness of PD with experimental and quasi-experimental designs rather than correlational studies

In PD research the majority of the quantitative studies are correlational surveys (Ball et al., 2008; Blank & de las Alas. 2009). In correlational survey studies, researchers observe and analyze natural variation in PD and student achievement without doing any manipulation (Wayne et al., 2008). In general these studies use large-scale survey data to examine the relationship between PD and student achievement. They are adequately powered and they are cost-effective. They produce knowledge about broad patterns of relationships between PD and student achievement but they provide limited information about details of the PD that are reported on by teachers (Seidel & Shavelson, 2007; Shavelson, & Towne, 2002; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). For example, Harris and Sass (2011) investigated the effect of number of contact hours of PD on student achievement by using Florida's state math and reading test data. In such studies, researchers generally try to adjust for the effect of ancillary factors (e.g., student and teacher demographics, school context) on student achievement by incorporating covariates in their statistical models. However, this method can't rule out alternative explanations for the observed effect of PD on student achievement (e.g., selection bias) as effective as random assignment of

the units to different PD conditions or matching PD and control conditions (Cook, 2002; Shadish et al., 2002; Shavelson, & Towne, 2002). Surveys used in these studies are also susceptible to measurement error since they collect broad information about PD experiences of teachers instead of asking teachers defined and discrete experiences that they have in specific PD program. This could introduce further bias in estimation. (Desimone, 2009; Mayer, 1999; Seidel & Shavelson, 2007) Hence, these studies are less likely to make causal claims about the effect of PD on student achievement (Cook, 2002; Wayne et al., 2008).

Different from correlational studies, in general, experimental/quasi-experimental studies examine the effect of a particular PD program on student achievement. They go beyond observational studies by including manipulation of the PD treatment. Thus, they are more likely to allow researchers to make causal inferences about the effectiveness of PD and to build a knowledgebase about what kinds of PD are effective for improving student achievement (Wayne et al., 2008). Moreover, Seidel and Shavelson (2007) found that in general experimental/quasi-experimental studies generated larger effect sizes compared to correlational studies. With the increasing incentives to conduct more rigorous PD studies, more experimental and quasi-experimental studies have been implemented recently (Blank & de las Alas, 2009; Raudenbush, 2008). Experimental/quasi experimental studies should be continued to increase the confidence we have in our inferences. In doing so, Seidel and Shavelson's (2007) review suggests we may also find more consistent effects of PD on student learning.

### 2.3.2 Experimental and quasi-experimental studies should be carefully designed in order to make strong causal claims

Recently, in the field of PD research, the number of randomized controlled trials (RCT), particularly group randomized controlled trials, has increased (Blank & de las Alas, 2009; Raudenbush, 2008). However, quasi-experimental studies in general have been preferred over RCTs because they are often more cost effective, and because they are more feasible to conduct in educational settings (Shavelson, & Towne, 2002).

From a statistical perspective, RCTs are considered the gold standard since random assignment of units to the conditions is the best way for ruling out all alternative explanations for the observed effect of the treatment (Cook, 2002; Shadish et al., 2002; Murnane & Willett, 2010).

However, RCTs do not necessarily always allow researchers to make strong causal claims. In contrast to laboratory settings, when RCTs are conducted in educational settings making causal claims becomes more difficult (McMillan, 2007). For example, in PD research, researchers have limited capacity to randomly assign teachers to the conditions (Shavelson, & Towne, 2002). Even when random assignment is possible; there is a high probability that the effect of the PD can spread to the control group since treatment group teachers interact with control group teachers within the same school (McMillan, 2007). As a result, most of the time schools are assigned randomly to the conditions instead of teachers (Borman, et al., 2008; Borman, et al.,2009; Garet et al., 2010, 2011; Matsumura et al., 2013; Newman et al., 2012). In that case, making causal claims about the effectiveness of PD requires keeping all factors but PD similar across treatment and control schools, which is also hard to accomplish. Thus using random assignment alone doesn't abdicate researchers of the responsibility of ensuring

equivalence of the treatment and control groups (McMillan, 2007). Researchers, therefore, need to carefully design experimental studies in order to make strong causal claims about the effectiveness of PD.

Quasi-experimental designs are preferred when random assignment is not feasible or not desirable (Shavelson, & Towne, 2002; Shadish et al., 2002). In quasi-experimental studies researchers have been aware of the fact that treatment and control groups might have pre-existing differences due to selection bias. Thus, they try to equate intervention and control group conditions to the extent possible. Some scholars argue that RCTs have many more potential threats undermining casual inference than highly controlled quasi-experiments (McMillan, 2007). Indeed, recent advances in the design and analysis of quasi-experiments (i.e. propensity score matching, regression discontinuity), allow researchers to minimize the degree to which selection bias contributes to alternative explanations for the observed effects (Shadish et al., 2002). These analytic methods can help researchers increase the confidence in attributing observed effects on student achievement to the PD itself (Wayne et al., 2008). Prior studies have typically not been able to employ these methods because these methods are relatively new and because these methods require not only vast amounts of available data, but considerable planning in the research design (Shadish et al., 2002). The degree to which quasi-experimental studies allow researchers to make causal claims depends on the extent to which the research design of quasi-experimental studies rule out competing explanations for gains in student achievement (Shadish et al., 2002; Murnane & Willett, 2010). Thus, in order to make strong claims about the effectiveness of PD, researchers should employ advanced techniques and carefully design quasi-experimental studies.

### 2.3.3 Data about fidelity of implementation of the PD should be collected

Like other educational interventions (e.g., comprehensive school reform interventions, curricular interventions, etc.), designs for PD research must account for the extent to which the PD is implemented with fidelity to the original design (Leinhardt, 1980; Stein et al., 2008). This is especially true when researchers attempt to understand a PD effort as it is scaled up beyond a single site (Berends, Kirby, Naftel, & McKelvey, 2001; Borko, 2004; Hamilton et al., 2003; Wayne et al., 2008). When PD programs are implemented at a single site, generally, educational researchers, scholars, or experts who also designed the PD program typically implement the PD. Thus, in such PD studies, it is high likely that PD[3] is delivered with fidelity to the design features of the PD (Borko, 2004; Wayne et al., 2008). However, when PD is conducted at scale, the PD program often reaches classroom teachers through a train-the-trainer approach. This approach requires that the PD providers not only learn new content, but also acquire new knowledge and skills for *helping others* to learn the new content (Stein, Smith, & Silver, 1999). Since it is hard for PD providers to adopt these new roles and acquire new skills, they may not implement PD consistent with its design features (Hamilton et al., 2003). Thus, in studies examining the effectiveness of PD, researchers need to collect data about the extent to which the PD is implemented with fidelity, especially in large-scale PD studies.

Without implementation data, we do not know whether the PD provided to teachers was consistent with the PD as it was originally conceived. For example, in a large-scale experimental study conducted by Borman et al. (2008), no data was collected about fidelity of implementation.

---

[3] For PD programs, either implemented in a single site or at large scale, there is always a risk for a decline in the level of fidelity due to the adverse effect of random factors (i.e. technical problems with equipment).

Thus we can not be sure whether the observed negative effect was due to the poor implementation of PD or due to the PD itself. If the PD was implemented with high fidelity, in cases where no effect was observed, implementation data would allow researchers to rule out one of the potential alternative explanations. When there is a positive effect of PD, knowing about the fidelity of implementation could provide more nuanced analytics about how PD produced a positive effect. This would add to researchers' confidence in the findings by limiting alternative explanations. Similarly, variation in implementation fidelity could also lead to variation in more proximal outcomes, such as classroom instruction. For example, Kisa and Correnti (2012) found that teachers in schools which were provided PD with high fidelity, exhibited different growth trajectories for reform aligned instruction compared to teachers in schools that provide PD with low fidelity to the reform ideals.

Data about fidelity of implementation of the PD can be collected in different levels of detail; from surface level (e.g. recording the duration of PD) to a more detailed level (e.g. observing/video-recording PD sessions to examine whether it is delivered consistent with the design principles). Collecting more detailed fidelity of implementation data requires more resources, human capital, time and strategic planning (e.g. how to code observations to assess level of fidelity). The more detailed the data, although it comes with a greater cost, the better researchers can answer questions about whether variations in treatment dosage lead to variations in the outcomes being measured.

Most large-scale experimental/quasi-experimental studies that I reviewed paid attention to the fidelity of implementation of the PD program (See Table 4). However, some studies collected more detailed data about how the PD was implemented (e.g. classroom observations) while others measured surface level aspect of fidelity of implementation (e.g. duration of PD).

Since fidelity of implementation data was collected primarily to confirm whether PD was implemented with fidelity or not, no variation in the level of fidelity of implementation was found, and hence it was not possible to explore a systematic relationship between the level of fidelity of implementation of PD and student achievement.

**Table 4.** Fidelity of implementation data collected in the recent experimental/quasi-experimental studies.

| Study | Sample | Component |
|---|---|---|
| Matsumura et al., (2013) | 167 teachers<br>29 schools | Self-reports |
| Harris et al., (2012) | 20 teachers* | - |
| Heller et al., (2012) | 270 teachers<br>6 states | Video-recording/observations<br>Attendance records |
| McMeeking et al., (2012) | 128 teachers<br>64 schools | Not collected |
| Biancarosa et al., (2010) | 287 teachers<br>17 schools | Not collected |
| Roth, et al., (2011) | 48 teachers* | - |
| Powell et al., (2010) | 89 teachers | Duration of coaching |
| Sailors & Price (2010) | 44 teachers<br>14 schools* | Video-recording/observations<br>Coaching reports |
| Garet et al., (2010, 2011) | 195 teachers<br>77 schools | Video-recording/observations<br>Attendance records,<br>Self-reports |
| Newman et al., (2012) | 475 teachers<br>82 schools | PD logs, Interviews<br>Self-reports |
| Matsumura et al., (2010) | 171 teachers<br>29 schools | Self-reports |
| Borman, et al., (2009) | 230 teachers<br>20 schools | Not collected |
| Heller (2012) | 181 teachers<br>137 schools | Video-recording/observations<br>Attendance records |
| Garet et al., (2008) | 270 teachers<br>90 schools | Video-recording/observations<br>Attendance records, Self-reports |
| Borman et al., (2008) | 80 schools | Not collected |
| * Indicates that PD was provided to teachers directly instead of using the train-the-trainer model. | | |

### 2.3.4 Effect of PD on teacher knowledge and practice should be examined

In order to improve student achievement, PD programs should first improve teacher knowledge and create change in teacher practice. The effect of PD on student achievement is thought to occur through mechanisms such as a change in teacher knowledge and classroom instruction. (Desimone, 2009). Effective PD programs designed to improve student learning, begin with an assumption that teachers need to acquire new knowledge about content, teaching strategies, student thinking, etc., and implement instructional practices that enable students to take more active roles in their learning in order to develop a rich understanding of important content (Borko & Putnam, 1995). Thus, in order to have a complete understanding of how PD creates change in student achievement, researcher should collect data on these proximal outcomes.

Moreover, there are fewer intervening factors between PD and teacher outcomes that could alter our determination of whether an effect of PD exists. If PD programs failed to create change in proximal outcomes, then researchers would need to find out why PD fell short in improving teacher knowledge and/or changing teacher practice. If, however, change in teacher knowledge and/or classroom practice was observed but no effect was found on student achievement, then other factors would need to be considered. Examining proximal outcomes provides information about *how* PD created gains in student achievement beyond simply informing *whether* an effect of PD on student achievement exists. In doing so, it helps researchers refute alternative explanations when an effect is present or could help identify where problems exist when no effect is observed on teacher or student outcomes.

In general, recent experimental/quasi-experimental PD studies have examined the effect of PD on teacher outcomes with varying approaches (See Table 5). Some of them have examined the effect of PD on teacher knowledge; some of them on teacher practice, and some of them have

examined the effect of PD on both outcomes. While one study used a teacher survey and another study used a student survey to measure instruction; the majority of the studies used classroom observations. To measure teacher knowledge, one study used a proxy measure (asking teachers about their knowledge), one study used a pure content knowledge assessment and the rest of them used measures of pedagogical content knowledge to examine proximal outcomes.

**Table 5.** Teacher knowledge and teacher practice data collected in the recent experimental/quasi-experimental studies.

| Study | Practice Data | Knowledge Data |
|---|---|---|
| Matsumura et al., (2013) | Observation | Not collected |
| Harris et al., (2012) | Not collected* | Not collected |
| Heller et al., (2012) | Not collected | PCK |
| McMeeking et al., (2012) | Not collected | Not collected |
| Biancarosa et al., (2010) | Not collected | Not collected |
| Roth, et al., (2011) | Observation | CK and PCK |
| Powell et al., (2010) | Observation | Not collected |
| Sailors & Price (2010) | Observation | Not collected |
| Garet et al., (2010, 2011) | Observation | CK and PCK |
| Newman et al., (2012) | Self-report | Self-report |
| Matsumura et al., (2010) | Self-report, observation | Not collected |
| Borman, et al., (2009) | Student reports | Not collected |
| Heller (2012) | Not collected | PCK |
| Garet et al., (2008) | Observation | CK |
| Borman et al., (2008) | Not collected | Not collected |

* Classroom observations were conducted just for providing feedback (not as an outcome).
CK=Content knowledge test, PCK=Pedagogical content knowledge test.

### 2.3.5 We should use high quality, proximal and student learning measures, which are sensitive to and aligned with the intervention.

In order to capture the effect of PD on student achievement, intermediate and student outcome measures should be well aligned with the focus of the PD. Assessing the effect of the PD on areas targeted by the PD would increase chances researchers would find an effect of the PD when one exists (Blank & de las Alas, 2009; Kennedy, 1998). For example, in a recent quasi-experimental study, Heller et al. (2012) could not detect an effect of three PD programs- Teaching Cases, Looking at Student Work, and Metacognitive Analysis- by using a selected response test as an outcome. However, when they examined the effects of the PD programs on students' written justifications they did detect differences between the effects of the three PD programs on student learning. Similarly, in another quasi-experimental study, Borman et al. (2009) found no effect of the PD on students' performances on a state test but did find a positive effect of the PD on students' knowledge that was measured by a more aligned test.

Review studies have also shown the importance of using specific and aligned outcome measures to assess the effectiveness of PD programs. They found that PD studies that used measures more aligned to the PD found larger effects compared to studies using more general state assessments (Kennedy, 1998; Blank and Las Alas, 2009). For example, Blank and Las Alas (2009) found that studies that utilized measures which were aligned to the focus of the PD (e.g., focus was teaching geometric concepts and students were assessed on knowledge of geometric concepts) had a mean effect size of .32. In contrast, the mean effect size was .01 for the studies that used statewide assessments in mathematics as an outcome measure.

Using well-aligned, specific measures as an outcome increases chances of finding an effect of the PD but there is an associated cost. Some measures such as having students write

extended text or conducting observations lack psychometric properties. In studies using such measures as an outcome, researchers need to ensure the quality of the measure (e.g. calculate inter-rater reliability/percent of agreement and provide validity evidence). Moreover, when the outcome measure is too aligned with the PD it may cover a limited range of students' knowledge or abilities and the scope of the generalizations of the results become very limited. On the other hand, state tests that are poorly aligned with the PD, but have established high psychometric properties (reliability, validity, objectivity) are less likely to capture effects if they are not aligned to the goals of the PD. However, if effects of PD on state test are found then this allows researcher to make broader generalizations for the effectiveness of the PD. Subject tests or benchmarks are in the middle ground in avoiding the pitfalls of these two measures. In general, they have established psychometric properties (reliability, validity, objectivity,) and also they are specific and aligned enough to the PD (Nitko, & Brookhart, 2011).

Alignment between outcome measures and the focus of the PD is also important for measuring intermediate effects of the PD. For example, in estimating the effect of PD on instruction, measures should represent the particular instructional practices that were the focus of content focused PD (e.g., Correnti, 2007; Matsumura et al., 2013). Similarly, teacher knowledge measures should also be aligned with the focus of the PD. Because the design of most PD programs includes a focus on pedagogical content knowledge in addition to content knowledge, measures of teacher knowledge should seek to include items aligned with both types of teacher knowledge (Kelcey & Phelps, 2013).

Recent experimental/quasi-experimental studies have used a variety of student outcomes (See Table 6). Some studies used the state test, several other studies used subject matter tests, and some studies used essay writing or written justifications as an outcome to measure the effect

of PD on student achievement. Studies that used essay writing or written justifications as an outcome generally reported inter-rater reliability and explained the scoring procedures but they couldn't provide validity evidence for the measures.

**Table 6.** Student outcome measures used in the recent experimental/quasi-experimental studies

| Study | Outcome Measure |
|---|---|
| Matsumura et al., (2013) | State Test |
| Harris et al., (2012) | Essay writing |
| Heller et al., (2012) | Electric circuits test<br>Written justification |
| McMeeking et al., (2012) | State Test |
| Biancarosa et al., (2010) | DIBELS (Basic Early Literacy Skills) subsets<br>Terra Nova |
| Roth, et al., (2011) | Science Content Knowledge Test |
| Powell et al., (2010) | Peabody Picture Vocabulary Test<br>Woodcock-Johnson-Letter Word Identification<br>Concepts About Print Measure<br>Alphabet knowledge Test<br>Test of Preschool Early Literacy-Blending<br>Initial sound matching measure and Writing |
| Sailors & Price (2010) | Group Reading Assessment and Diagnostic Evaluation |
| Garet et al., (2010, 2011) | Rational number test |
| Newman et al., (2012) | SAT-Problem solving subtest<br>SAT Science |
| Matsumura et al., (2010) | State test<br>DRP Degrees of Reading Power Assessment |
| Borman, et al., (2009) | Subject matter test<br>State test |
| Heller (2012) | Force and motion test |
| Garet et al., (2008) | Terra Nova (In four sites)<br>SAT-10 (In one site)<br>A criterion reference test (In one site) |
| Borman et al., (2008) | Life science tests, Earth science tests<br>Physical science tests |

### 2.3.6 We should examine effects of PD longitudinally

It is difficult to separate students' status from their growth in education research (Rowan, Correnti, & Miller, 2002). Yet, we are more interested in the effect of PD on students' growth in achievement (Bryk & Raudenbush, 1987; Betebenner & Linn, 2010). In order to measure growth, studies need to collect student achievement for at least three time points (Bryk, & Raudenbush, 1987). Studies that measure student achievement at one time point (i.e. cross-sectional studies) or at two time points (covariate adjusted models) have two problems. First, they attempt to estimate effects of PD on students' status (this is true even when the models employ a covariate adjustment for prior achievement, as demonstrated in Rowan et al., 2002). Second, they frequently attempt to observe the effect on students during the same year the PD was being implemented. Compared to these models, using statistical models that directly estimate students' individual growth trajectories such as growth modeling, can more properly estimate the effect of PD on change in student achievement (Bryk & Raudenbush, 1987). In fact, as Rowan et al. (2002) have shown, effect sizes estimated in growth model about two-to-three times larger than what they calculated using a simple covariate adjustment model (growth models: Cohen's d = .72 to .85; covariate adjustment models: d = .21 to .42).

Moreover, by using growth models, we can estimate the effect of PD on both achievement growth and achievement status. For example, by employing a growth model Powell and his colleagues (2010) found positive effects of remote and onsite coaching both on growth rate and mean scores of preschool children's letter knowledge, blending skills, writing, and concepts about print. Longitudinal designs could also allow researchers to examine whether the effects of PD on student achievement sustain, accelerate or fade over time (Tan, 2008). For

example, a growth model employed by Borman et al. (2009), demonstrated that at the end of the first year of the PD, students whose teachers participated in the PD performed similar to students whose teachers did not participate to the PD. However, over multiple years they demonstrated that students of intervention teachers increased their content knowledge more than their peers. Finally, growth models also provide information about the trend of the effect along with more refined judgments about whether the PD improved student achievement or not.

Table 7 below shows that most recent experimental/quasi-experimental studies have moved beyond cross-sectional designs because of the inherent weakness in measuring outcomes at a single time point. However, the majority of them used covariate-adjusted models to measure the effect of PD on student achievement. There are only a small number of studies that have employed some sort of growth modeling to examine the effect of PD on both the status and growth of student achievement.

**Table 7.** Statistical models employed in the recent experimental/quasi-experimental studies.

| Study | Statistical Model |
|---|---|
| Matsumura et al., (2013) | Covariate adjusted Moderation model |
| Harris et al., (2012) | Covariate adjusted HLM |
| Heller et al., (2012) | Covariate adjusted HLM |
| McMeeking et al., (2012) | A logistic, generalized linear mixed model |
| Biancarosa et al., (2010) | Accelerated Longitudinal Cohort Model |
| Roth, et al., (2011) | Growth Model HLM |
| Powell et al., (2010) | Growth Model HLM |
| Sailors & Price (2010) | Covariate adjusted HLM |
| Garet et al., (2010, 2011) | Covariate adjusted HLM |
| Newman et al., (2012) | Covariate adjusted HLM |
| Matsumura et al., (2010) | Covariate adjusted HLM |
| Borman, et al., (2009) | Growth Model HLM |
| Heller (2012) | Covariate adjusted HLM |
| Garet et al., (2008) | Covariate adjusted HLM |
| Borman et al., (2008) | Covariate adjusted HLM |

### 2.3.7 Populations of students taught by treated teachers should be as similar to students of comparison teachers as possible

It is challenging to isolate effects of PD from other factors such as the curriculum experienced by students (Scher & O'Reilly, 2009). When attempting to contrast outcomes from treated and control classrooms, if teachers were implementing different curricula in their classrooms this would create additional variation between classrooms in students' achievement in addition to whatever variation was caused by the PD. In such cases, attributing an observed effect on student achievement to the PD alone would not be possible (third condition of causation). In designing

PD studies educational researchers should try to account for these important confounds by keeping the PD and control conditions as similar as possible. For example, insuring a study has both treated and comparison teachers implementing the same curriculum would help researchers control one key factor thought to influence student achievement (Ball & Cohen, 1996). In cases where it is not possible to do this by design, researchers should collect data about these varying factors and try to adjust for them in their analytic models. Identifying a smaller group of teachers in both PD and control groups with similar values on these factors and matching these groups can help researchers to control effects of these factors. Matching teachers, as much as possible, can help rule out alternative explanations (Murnane & Willett, 2010; Shadish et al., 2002; Wayne et al., 2008). No single study will be able to control all conditions, but to the extent possible researchers should try to limit large structural differences between groups to help reduce the extent to which inferences are susceptible to such forms of selection bias.

Some of the experimental/quasi-experimental studies that I reviewed were able to keep the curriculum similar across conditions to ensure that the observed effect of the PD was not confounded with the curriculum. Meanwhile, other studies could not keep the curricula similar across conditions (See Table 8). Moreover, in some studies PD was designed to support teachers' implementation of the new curriculum. In such studies, separating the effect of two treatments (PD and curriculum) could be an issue. However, by offering the new curriculum to the control group teachers as well, Borman et al. (2008) and Newman et al. (2012) could separate the curriculum's effect from the effect of the PD. In only one of the experimental/quasi-experimental studies I reviewed (Borman et al., 2009), it was not possible to isolate the PD's effect from the effect of the curriculum since the new curriculum was not offered to teachers in the business-as-usual condition.

**Table 8.** Similarity of the curriculum across treatment and control conditions in the recent experimental/quasi-experimental studies.

| Study | Curriculum across conditions |
|---|---|
| Matsumura et al., (2013) | Similar |
| Harris et al., (2012) | No Information available |
| Heller et al., (2012) | Not similar |
| McMeeking et al., (2012) | No Information available |
| Biancarosa et al., (2010) | No Information available |
| Roth, et al., (2011) | Not similar |
| Powell et al., (2010) | Not similar |
| Sailors & Price (2010) | No Information available |
| Garet et al., (2010, 2011) | Not similar |
| Newman et al., (2012) | Similar |
| Matsumura et al., (2010) | Similar |
| Borman, et al., (2009) | Not similar |
| Heller (2012) | Not similar |
| Garet et al., (2008) | Not similar |
| Borman et al., (2008) | Similar |

Each of these elements of research design can help to make stronger inferences by limiting alternative explanations and to find more consistent effects of PD on student learning. However, incorporating all of these elements in an effective manner (rather than superficially) into research designs is challenging due to the complex nature of educational settings. Moreover, how researchers address these design features involves trade-offs (e.g. level of alignment of student outcome with the PD) because some design elements (e.g. collecting observation data) require more resources, human capital, and time. As a result, even recent experimental/quasi experimental studies, which are more rigorous than the majority of previous quantitative PD

studies, could only incorporate some of these design features and they employed them with varying levels of quality.

Identifying patterns in how these studies incorporated these design principles combined with insights about the significance of the PD effect found in these studies is difficult because there are too many changing variables (e.g., PD studies have integrated different design elements with varying quality and they also have different duration/contact hours). However, it stands to reason that the more studies attend to these principles, the more likely they will be able to find consistent effects of PD (i.e., when there is an effect). Of course this also depends on the extent to which the PD programs being investigated are designed with care.

## 2.4 PREREQUISITE FOR DESIGNING EFFECTIVE PD RESEARCH

PD programs examined in recent experimental/quasi experimental studies were content focused, relatively intense and of longer duration. However, contrary to what is observed in these studies, national surveys have continually indicated that large numbers of teachers still attend brief, one-shot workshops for their PD (Birman, Desimone, Garet, Porter, & Yoon, 2001; Darling-Hammond et al., 2009; Hill, 2007; Parsad, Lewis, & Farris, 2001). These workshops typically provide piecemeal instructional activities, and the design of this type of PD is widely thought to be ineffective in creating changes in teachers' practice and student achievement (Ball & Cohen, 1999; Ball & Forzani, 2009; Birman et al., 2001; Corcoran, 1995; Darling-Hammond et al., 2009; Hawley & Valli, 1999; Hill, 2007; Parsad et al., 2001; Stein, et al., 1999). Considering this, it is important to emphasize that in order to create change in teacher knowledge and practice, and improve student achievement, PD should be content focused, intense and

implemented for a longer duration (Blank & de las Alas, 2009; Clewell et al., 2004; Kennedy, 1998; Scher & O'Reilly, 2009; Yoon et al., 2007). Recommendations for better design of PD studies can only meet with success when the PD design itself has these features. If the PD that is the focus of study is not content focused, for example, assessing its effectiveness with well-designed research would not help finding any effect of PD.

## 2.5 CONCLUSION FOR LITERATURE REVIEW

This review of the literature has revealed that there are a proportionally small number of experimental/quasi-experimental studies that have examined the effectiveness of PD for improving student achievement across all subject areas. These studies have produced mixed evidence for the overall effectiveness of PD and for specific features that are thought to be instrumental for making them effective. In PD research, there are many factors that could have a role in the observed effect of PD such as low fidelity of PD implementation, low alignment between outcome measures and the focus of the PD, effectiveness of the curriculum, etc. Most of the PD studies, including some high quality experimental/quasi-experimental studies, ignored the role of these factors on the observed effect of PD, or could only control their influence to a limited extent due to the complex nature of educational settings. This limited control over the factors that potentially confound observed effects of the PD prevents researchers from making stronger inferences about the effectiveness of PD and finding more consistent effects of PD on student learning.

In order to build our knowledge base about the effectiveness of PD for improving students' achievement in mathematics (and in all subject areas), we need to implement more

carefully designed experimental/quasi-experimental studies. Such studies should attend to the fidelity of implementation of PD; examine intermediate teacher outcomes of PD; use proximal and student learning measures which are sensitive to and aligned with the PD; examine longitudinal as well as cross-sectional data and potentially employ growth modeling; and try hard to make populations of students taught by treated teachers as similar to students of comparison teachers as possible.

# 3.0 CONTEXT OF THE STUDY

This study produced empirical evidence for the effectiveness of an ongoing, content-focused PD program for students' mathematics achievement. The PD program was designed and implemented as a part of a Math and Science Partnership (MSP) grant program with a collaborative effort of a non-profit organization, a university and an urban school district located in the Northeastern U.S.

## 3.1 GOALS AND FEATURES OF THE MSP-PD

The ultimate goal of the MSP-PD was improving the mathematics learning of all students. The PD aimed to immerse mathematics teachers as learners in learning environment that they expected to create for their students in which students would engage in doing rigorous, connected and interesting mathematics; mathematics that reflects the discipline (Cuoco, Goldenberg, & Mark, 1996; Burton, 2004) Such a learning environment helps students deeply develop and make sense of mathematical practices and empower them as doers of mathematics in order to improve their learning and achievement of mathematics, (Kilpatrick, Swafford, & Findell, 2001). In order to create such learning environments, the MSP-PD aimed for teachers to acquire a disciplinary, effort-based approach of teaching mathematics. The goal of the PD was to be able to meet students where they are and help them coherently develop a profound

understanding of mathematics and the myriad mathematical ideas, tools, and skills which are necessary to be a doer of mathematics.

The MSP-PD provided coherent, sustained, and intense learning opportunities for teachers in order to provide them a model and make them reflect on what constitutes an effective mathematics learning environment in their classrooms. One of the foci of the PD was helping teachers to effectively implement the CME geometry curriculum. CME is a student-centered, standards-based curriculum that emphasizes students' developing the ways of mathematical thinking-the habits of mind-used to create mathematical results by engaging them in the process of creating, inventing, conjecturing, and experimenting. Since this approach to teaching mathematics was new to teachers, engaging teachers as learners with the instructional activities that they were required was one of the critical design features of the PD.

Another key feature of the PD was focusing equally on both teachers' content knowledge of mathematics and specialized knowledge of how to teach the content effectively. Through focusing on developing mathematical practices, the PD included teachers working on content as mathematicians and engaging teachers in mathematics as a discipline, to better prepare them to see and hear the mathematics in students' work (Cuoco, 2001). To improve teachers' specialized content knowledge for effective teaching, MSP-PD provided opportunities for teachers to reflect on their own disciplinary experiences as learners and its implications for creating effective learning opportunities for their students.

## 3.2 PRACTICES OF THE MSP-PD

The MSP-PD was provided by an expert mathematician from a not-for-profit organization and from district professional development staff. During the summer institute an expert mathematician took the lead with the district's PD staff supporting instruction. Follow-up sessions during the year were conducted by the district's PD staff.

In the first week of the summer institute, MSP-PD teachers worked as a group on purposefully constructed inquiry-based problem sets designed to focus teachers on connections between geometric ideas and algebra1 through hands-on learning activities. PD providers mentored, facilitated and guided teachers while they were working on the problem sets. The problem sets along with the PD providers' guidance, allowed teachers to do math, to build their own disciplinary experiences, to explore deep mathematical ideas and to develop learner experience around geometry and algebra1. Moreover, this learning experience helped them to further improve their mathematical thinking skills and supported their abilities to understand the mathematical thinking of students.

In week 2, teachers worked on the CME geometry curriculum and how to apply the habits of mind approach while teaching the curriculum. Teachers ran some of the activities as if they were students while, at the same time discussing the purpose of those activities. Teachers reviewed chunks of the curriculum and analyzed how and for what purpose topics are introduced in the specific order. PD providers helped teachers make connections between their own learning and disciplinary experiences that they had during the first week of the PD while teachers were reflecting and discussing the implications of the new curriculum (habits of mind) for their instruction and how to teach inquiry-based CME geometry curricula effectively.

During the school year MSP-PD teachers were provided 6 follow up sessions every three weeks on Saturdays. Follow up sessions started at the end of September and ended at the beginning of February. Each follow up session was 4 to 5-hours long. During the follow up sessions, teachers focused on applying their learning to their instruction by collaboratively analysing the big ideas of the new geometry curriculum. Moreover they developed mathematical projects to support student facility with standards for mathematical practice. Either individually or in pairs, they investigated a mathematical question of their choice. They produced write-ups and a research report while working on the project. This experience helped them to develop autonomy to ask and investigate mathematical questions, which were deemed a critical skill for teachers as doers of mathematics. In doing so, they also learned to support students to develop their own autonomy.

# 4.0 METHOD

## 4.1 OVERVIEW OF THE RESEARCH DESIGN

In order to precisely examine the effectiveness of the MSP-PD for improving student achievement, I incorporated some of the design elements for effective PD research suggested in the literature review. For example, I employed a quasi-experimental research design in order to generate robust estimates of MSP-PD effects. To better detect the effectiveness of MSP-PD on changes in student achievement over time, I used growth models to examine the effect of the MSP-PD and contrast students in classrooms of MSP teachers versus matched comparison classrooms. These models accounted for the nested structure of the data (time points were nested within students, students nested within classrooms) while also allowing the ability to adjust for the effects of student background characteristics in the models. I also examined the effect of MSP-PD on teacher content knowledge to contribute further evidence that the MSP-PD was effective in producing change not only for distal outcomes, but for a proximal outcome as well.

In this study, treatment and comparison group teachers were teaching in the same school district. This helped me keep MSP-PD and comparison classrooms as similar as possible since the structure of each mathematics course I examined was the same across schools within the district. This design feature was also critical for ruling out some important competing explanations for gains in student achievement. Because of the district policy regarding how

students were tracked in their math courses, student populations in classrooms of the same course were similar in terms of students' prior mathematics achievement. Thus, I ran analyses separately for each course in order to ensure populations of classrooms taught by MSP-PD and comparison group teachers were similar across classrooms. At the same time, by looking within courses, the effect of the curriculum on the observed effect of MSP-PD, one of the key factors potentially influencing student achievement, was controlled because MSP-PD teachers and comparison teachers were implementing the same curriculum. Moreover, I used curriculum-based assessments (CBA) as a student outcome measure. CBA assessments were formed by the district to measure students' performance relative to each course in the curriculum. Thus, CBAs helped to create a coherent system, where the MSP-PD, curriculum, and assessments were all aligned. In addition to increasing researchers' chances of finding an effect of MSP-PD, the coherence of the system itself is thought to have potentially increased the effectiveness of the MSP-PD, also making it more likely to uncover effects on student achievement.

Examining the effect of MSP-PD within a given course also ensured the curriculum and the assessment was exactly the same for our contrasted groups of students. The student populations were also similar across MSP-PD and comparison classrooms. Yet, there might still be differences between MSP-PD and comparison classrooms due to teachers volunteering for the treatment. To address this, I matched classrooms taught by MSP-PD and comparison group teachers by using propensity score stratification methods. Given that I was making between classroom comparisons, I matched classrooms based on the classroom and school contexts and then checked that my treatment and comparison groups were matched on all pre-treatment covariates (n=54). This helped me to control for possible pre-existing differences that might have confounded the observed effects of MSP-PD on student achievement.

Technically, in a single case study it is not possible to test or determine the extent to which incorporating these design features was instrumental for finding an effect of PD on student achievement. Simply put there is no way to understand the counterfactual condition: Once the study is designed, re-designing the same study in a different way becomes impossible. While this study doesn't claim to empirically test the effectiveness of the research design features it is important to reflect on the importance of well-designed research for building the knowledge-base of whether and how PD can have effects on teaching and learning. By employing certain design features in this study, confounding factors such as curriculum and students' ability were controlled and that, in turn, increased my confidence in attributing any observed effects on student achievement (including null effect) to the MSP-PD itself.

## 4.2 SAMPLE

Algebra1, geometry, and algebra2 teachers in an urban district located in the Northeastern United States who voluntarily attended the MSP-PD formed the treatment group for this study. Teachers in the district who didn't attend the PD formed the comparison group.

In the sample of the study, there were two courses for each subject: tier1 and tier2. Tier1 courses were offered for students who had higher prior mathematics achievement while tier2 courses were offered for students who had lower prior mathematics achievement. For analyzing the effects of MSP-PD on student achievement, I ran a growth model specific to tier1 and tier2 courses within each topic (i.e., algebra1, geometry, algebra2). Teacher and student samples for

tier1 and tier2 courses of each topic were as follows: In algebra1, 13 tier1 teachers[4] and 17 tier2 teachers: in geometry, 14 tier1 and 7 tier2 teachers: in algebra2, 16 tier1 and 10 tier2 teachers formed the sample. Across tier1 and tier2 courses within each subject, there were in total 847 students in algebra1, 1060 students in geometry, and 1068 students in algebra2 classrooms. In total, 2975 students from 77 classrooms taught by 65 high school mathematics teachers formed the sample of the study. Number of classrooms and students in MSP-PD and comparison groups for each course were provided in Table 9. Analyses examining the effect of MSP-PD on teachers' content knowledge were done using data from 29 teachers who attended the MSP-PD (regardless of the content of the courses they taught). 15 teachers who volunteered to take the content knowledge assessment provided comparison group data.

**Table 9.** Number of classrooms and students in MSP-PD and comparison groups for each course.

| | Tier1 | | Tier2 | |
|---|---|---|---|---|
| | MSP-PD | Comparison | MSP-PD | Comparison |
| **Algebra1** | | | | |
| Classrooms | 6 | 7 | 7 | 10 |
| Students | 177 | 210 | 129 | 331 |
| **Geometry** | | | | |
| Classrooms | 6 | 8 | 3 | 4 |
| Students | 336 | 479 | 105 | 140 |
| **Algebra2** | | | | |
| Classrooms | 7 | 9 | 3 | 7 |
| Students | 327 | 448 | 108 | 185 |

[4] Unit of analysis within each course is the classroom(s) taught by an individual teacher. This is as close to individual classrooms as I could get based on the data provided by the district. Same teacher was sampled in different models if he taught different courses within a school year.

MSP-PD teachers attended the two-week long summer institute and also the six follow-up sessions. Comparison group teachers received a one-day kick-off PD that was provided by the district to all teachers (including MSP-PD teachers). Since the new geometry curriculum was introduced to teachers, in addition to kick-off day PD, comparison group geometry teachers were provided PD to introduce them to the new geometry curriculum. In 5 or 6 half-day sessions this PD covered the big ideas of the curriculum, how ideas grow, and also engaged teachers with some practices within the curriculum. Even though their attendance was optional some MSP-PD teachers may also have attended this PD.

## 4.3 OUTCOME MEASURES

### 4.3.1 Student Outcome

In this study, I examined students' achievement trajectories within an academic year by employing separate longitudinal models for each mathematics course. These models compared math achievement trajectories of students whose teachers attended MSP-PD versus students whose teachers didn't attend the MSP-PD. I used Curriculum Based Assessments (CBA) as an outcome in these models.

CBA assessments were criterion-referenced assessments developed by the district. They contained both constructed-response and multiple-choice items. The CBA assessments were used to measure student learning of the district's mathematics curriculum and to inform instruction to help monitor students' progress during the school year toward meeting the standards. They were administered three times within the school year in November, January, and June. All students

were administered the CBAs at the same time. Since new concepts and skills in the math curriculum build upon previously taught concepts and skills, parallel with the curriculum, skills and concepts measured in CBAs administered towards the end of the year built on the skills and concepts measured in previous administrations of the CBAs (See Appendix A). Moreover, consecutively administered CBAs included some common items measuring the same concepts for some question types (e.g., the concept of slope in algebra1).

CBAs for each subject were the same for both tier1 and tier2 courses except for algebra1. For algebra1, there was a specific CBA for tier1 courses and a slightly different one for tier2 courses. The content was roughly the same across the entire year, however, the order of topics and therefore the order of items was slightly different for tier1 and tier2 algebra1 courses because the assessments were aligned with the order of topics in the curriculum.

### 4.3.2 Teacher Outcome

The Knowledge of Algebra Teaching (KAT) assessment was used in order to measure the effect of PD on teachers' content knowledge. The KAT was produced by a group of researchers at Michigan State University for measuring teachers' pedagogical content knowledge of algebra at the secondary school level as part of the project Knowledge of Algebra for Teaching (See Ferrini-Mundy, Burrill, Floden, & Sandow, 2003; Ferrini-Mundy, McCrory, & Senk, 2005; Floden & McCrory, 2007). KAT was designed to assess three aspects of teachers' content knowledge; 1) knowledge of school algebra—how well the teacher can solve problems at the middle and high school level; 2) advanced knowledge of algebra—how well the teacher understands college-level algebra (e.g., calculus, abstract algebra); 3) teaching knowledge—how well the teacher understands the challenges that students might have with particular algebraic

concepts and skills, including knowledge of typical student errors, canonical uses of school math, curriculum trajectories, etc. Thus, KAT was well aligned with the MSP-PD focusing on teachers' content knowledge of mathematics and knowledge of how to teach mathematics effectively. As far as content is concerned, KAT and MSP-PD were also well aligned given that central focus of the MSP-PD on algebra along with geometry.

MSP-PD teachers took KAT Form A in August, prior to the MSP-PD intervention. Right after the MSP-PD summer workshop, they took KAT Form B (alternate form), as a post-test. Thus, analysis examining the effect of MSP-PD on teachers' mathematical content knowledge, in fact, only examined the effect of the MSP-PD workshops not the whole MSP-PD including follow up sessions. Comparison group teachers took the KAT Form A only one time in the late Fall term just after their recruitment.

## 4.4 STATISTICAL ANALYSIS AND MODELS

### 4.4.1 Analyses examining the effect of MSP-PD on teachers' mathematical content knowledge

In order to understand whether MSP-PD teachers grew in their content knowledge, I examined whether MSP-PD teachers' KAT scores increased significantly in the post-test by comparing their post-test and pre-test scores. I used a paired-samples t-test for this comparison. Next, in order to understand how different the treated teachers were from the overall population of high school mathematics teachers in the district, I compared the mean pre-test scores of the MSP-PD

teachers versus the comparison group teachers' KAT scores. For these analyses, I employed an independent-samples t-test.

### 4.4.2 Analyses examining the effect of MSP-PD on students' mathematics achievement

I examined the effect of MSP-PD on students' achievement by employing separate repeated measures models for each course. The level of confidence in attributing observed effects detected in these models to the MSP-PD depends on how effectively the research design ruled out alternative explanations for the effect. Thus it was important to create comparable classrooms, which were different only by being taught by MSP-PD and non-MSP-PD teachers. In order to create comparable classrooms, I examined longitudinal models at the course level (e.g., algebra1 tier1) and performed propensity score stratification to match classrooms on their aggregate classroom characteristics and school contexts within courses.

Although teachers were provided the MSP-PD, a strategic decision was made to compare classrooms taught by teachers attending the MSP-PD versus a set of comparison classrooms as similar as possible to these classrooms in all other ways. Initial analyses demonstrated the futility in comparing MSP-PD teachers with comparison group teachers because they taught different topics in different configurations to different students of differing abilities and at different grade levels. Thus, matching by teacher did not make for very close matches between MSP-PD teachers and comparison group teachers. Instead, by comparing classrooms, it was possible to create close matches between classrooms within the same course, with students taking the exact same curriculum and assessments. Furthermore, because of the district tracking policy, the classrooms of MSP-PD and comparison group teachers, within courses, were quite similar from the outset of our matching. To make classrooms even more similar and adjust for possible

differences in school environments, I employed propensity score stratification methods to match classrooms taught by MSP-PD teachers with classrooms of the same course taught by comparison group teachers.

Moreover, running models at the course level also allowed me to compare and contrast across courses and types of students served by each course. For example, I could explore whether the MSP-PD served better to improve students' achievement in algebra1 and geometry (relative to algebra2) because of the content focus of the PD. Results of this comparison could reveal whether alignment between focus of the PD and the measured outcome(s) matter for identifying an effect of the MSP-PD. Similarly, given the district's allotment of students to mathematics courses, I was also able to explore whether the effect of the MSP-PD was more influential (or, alternately, more easily detected) for the achievement of low-performing students than it was for high-performing students.

### 4.4.3 Propensity score stratification method

Random assignment of teachers to the conditions was not feasible, however, by matching MSP-PD and comparison group classrooms, this study simulated an experimental study that I would have implemented if teachers had been able to be randomly assigned to the conditions. Since teachers volunteered for the MSP-PD, there might be pre-existing differences between MSP-PD and comparison group teachers in terms of student, teacher or school related factors. To address these possible differences, I matched classrooms taught by MSP-PD and comparison group teachers by using propensity score stratification methods.

By employing propensity score stratification methods, a researcher can estimate each subject's probability (propensity) of being in the treatment group as a function of all observed

covariates. Then, a researcher can create subsets (strata) of treatment and control group subjects with similar estimated treatment propensity scores. Since the estimated propensity score can be thought of as a variable that is a summary of the set of covariates from which the propensity score is estimated, within each stratum, the distribution of the observed covariates would be the same across treatment and control conditions (Rosenbaum & Rubin, 1983). Therefore, stratifying on the estimated propensity score is expected to remove selection bias due to the effect of observed covariates under the assumption of strongly ignorable treatment assignment. Thus, by pooling estimates within these strata, a researcher then can estimate the average causal effect of a treatment on the outcome.

Propensity score stratification methods include several steps. In order to generate matches between classrooms within courses (i.e., tier1 and tier2 courses of algebra1, algebra2, and geometry topics), I followed the same steps. As a first step, I selected a set of observed covariates from which to estimate the propensity score. This step was critical for meeting the assumption of strongly ignorable treatment assignment. This assumption states that unobserved covariates are unrelated to treatment assignment given that all relevant covariates have been controlled for. The degree that I adjust for selection bias through matching and hence the strength of the inferences depends on the assumption that the observed covariates that I selected are more likely to confound treatment than any unobserved covariates. I selected a large set of observed covariates (54 covariates), which might theoretically confound the treatment to estimate the propensity score. These covariates include pretreatment characteristics of teachers, demographic characteristics of the schools that they were teaching in, and demographic characteristics of their classrooms (See Appendix B).

In the next step, I created a propensity score measure for each classroom. The propensity score in this case is the estimated probability of the classroom being taught by a treated teacher. The propensity score for each classroom was generated using a logistic regression model as a function of pretreatment covariates. It is known that when logistic regression models are run with small data sets the models might provide overconfident estimations. Higher estimations might be estimated too high, and low estimations might be estimated too low (Steyerberg, Eijkemans, & Habbema, 2001). In my analysis, the number of classrooms within courses was relatively few. Thus, I used Penalized Maximum Likelihood Estimation (PMLE) to adjust for over-optimism in propensity score estimations. I used the 'rms' package in the statistical program R (Harrell, 2013). It provided me an assessment of the model's degree of over-optimism and generated suggested penalty factor with the highest Akaike Information Criterion (AIC) for the adjustment. I applied this penalty factor to logistic regression in order to obtain the adjusted propensity scores.

After generating the propensity score, the next step was to create strata of classrooms by grouping classrooms that had similar estimated propensity score together. Several conditions should be achieved simultaneously to create strata for statistical analysis. First, classrooms of MSP-PD and comparison group teachers should be balanced (should have similar distribution) on each of the 54 observed pretreatment covariates. Second, means between the estimated propensity scores of the classrooms taught by MSP-PD and comparison group teachers within each stratum should be similar. Lastly, in order to generate MSP-PD effect within stratum, each stratum should include at least one classroom taught by MSP-PD and at least one classroom taught by a comparison group teacher. In cases when any of these conditions wasn't met, I modified the logistic regression equation, adjusted for over-optimism, re-created strata and

checked these conditions again. After several iterations of these steps, I achieved strata meeting these conditions for each course.

Within each course, I created two strata by using a median split after ranking classrooms based on their adjusted propensity scores. The logistic regression models that generated these propensity scores included only school level variables as a predictor. As previously described, classrooms were already very similar in terms of student population, due to district policy about student placement into courses. All strata were balanced on the 54 observed pretreatment covariates. Using significant testing at a p-value of .05, for none of the courses, proportion of the significant results was more than 5% of all mean and proportion comparisons[5]. These results indicated that with 2 strata that were created using median propensity score, balance was achieved across the covariates.

The final step in my statistical analyses was using dummy coded strata variables in associated growth models to adjust for selection bias as a result of volunteerism. I entered one stratum in each model in order to leave the other one out as a reference.

### 4.4.4 Growth Models

In order to answer my research questions concerning the effect of the MSP-PD on students' mathematics achievement, I examined three-level hierarchical linear repeated measures models (HLM) (Raudenbush & Bryk, 2002) in which the CBA assessment at each time point was nested within students and students were nested within classrooms.

[5] I reached this conclusion considering the fact that with large samples and random assignment 5% of the comparisons would be shown to have significant results by chance by the definition of p-value 0.05. This result indicated that balance was achieved at a level even better than chance (i.e. greater than 95% of the t-tests/z-tests are non-significant).

In these models students' individual achievement trajectories of standardized percent correct on the CBA assessments comprised level-1 of the model. Using standardized scores at each administration of the test adjusted for the test difficulty while also equating scores for each CBA administration[6]. Time was centered at the last time point (i.e., the June administration) in order to understand the differences between students in their status at the end of the year as well as their change in achievement trajectories over the course of the year. I entered student level covariates at level 2. Although aggregates of these variables were used to calculate propensity scores for matching classrooms taught by MSP-PD and control group teachers, I entered them into the model at their original units as well, in order to account for any possible information lost during aggregation. This in fact, increased precision of the estimated effect of the MSP-PD by reducing the standard error associated with the effect. The general form for these models is provided below. While the general form for the outcome algebra1 tier1 is given, the model is exactly the same for all courses.

Level 1 Model (Time Points)

$$(Alg1\ Tier1)_{tij} = \pi_{0ij} + \pi_{1ij}(Time)_{tij} + e_{tij}$$

[6] When the raw score for each student was used at each time point, fluctuations in the difficulty of the assessment can be falsely reflected as an increase or decline in students' achievement in the growth models. For example, consider a student who gets 50 percent of the test correct on the first test (while the average percent correct is also 50 percent), and then gets 75 percent of the test correct on the second test (while the average percent correct is 80). Using the raw percent correct would result in the appearance of gains for this student, but his/her performance relative to others actually decreased on the second test. Standardizing the outcome at each time point provides a relative measure of performance at each time point – where students at each time point are compared on a standardized scale with the performance of all students taking the CBA at that time point. Thus, growth measures are relative to the other students taking the same assessments. This comparative measure allows us to draw an inference about the growth of one group versus another.

Level 2 Model (Students)

$$\pi_{0ij} = \beta_{00j} + \sum_{q=1}^{Q_p} \beta_{0qj} X_{ij} + r_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \sum_{q=1}^{Q_p} \beta_{1qj} X_{ij} + r_{1ij}$$

Level 3 Model (Classroom)

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(MSP-PD)_j + \gamma_{002}(Stratum)_j + u_{00},$$

$$\beta_{10j} = \gamma_{100} + \gamma_{100}(MSP-PD)_j + \gamma_{100}(Stratum)_j + u_{10},$$

$$\beta_{0qj} = \gamma_{0q0}$$

$$\beta_{1qj} = \gamma_{1q0}$$

Estimates from these models allowed me to examine the effect of MSP-PD on both the growth rate and status of student achievement for each course. The growth rate for MSP-PD ($\gamma_{101}$) indicated how the standardized mean performance of students in classrooms taught by MSP-PD changed on the CBA assessments over time relative to the performance of comparison teachers' students. Status for MSP-PD ($\gamma_{100}$) indicated how students in classrooms of MSP-PD teachers performed compared to students of comparison group teachers at the end of the school year. Both estimates were calculated after having matched treated and comparison classrooms and adjusting for student differences within classrooms.

Lastly, since all six outcome measures were standardized and the same covariates were used across the models, effect sizes of MSP-PD for students' growth rate and status were comparable. This allowed me to explore whether the estimated effect of the MSP-PD varied depending on the alignment between PD content and student outcomes being investigated and

whether it varied based on differences between student populations in terms of their prior math performances.

# 5.0 RESULTS

I present results of the analyses in two stages. The first stage examines the effect of MSP-PD on teachers' mathematical content knowledge. I hypothesized that content focused and intense MSP-PD helped participant teachers improve their mathematical content knowledge. I conducted t-test mean comparisons to examine this hypothesis. In the second stage of my results, I present results from the analyses that were the main focus of this study. I hypothesized that attending effective PD generated improvements in teaching for MSP participants leading to improvement in their students' math achievement. I describe results from these HLM growth models examining the effect of MSP-PD on student achievement. In these models, I compared the growth trajectories of students whose teachers attended the MSP-PD relative to matched classrooms of students whose teachers didn't attend the MSP-PD.

## 5.1 EFFECT OF MSP-PD ON TEACHERS' MATHEMATICAL CONTENT KNOWLEDGE

In order to examine the effect of MSP-PD on teachers' content knowledge, I first examined whether any improvement occurred within the group of teachers participating in the MSP-PD. Because the teachers were administered the post-test KAT just after the MSP-PD workshop, inferences about the observed effects are confined to the effects of the workshop only.

Results of the paired-samples t-test revealed that the MSP-PD teachers' mathematical content knowledge of algebra was significantly higher after attending the MSP-PD workshop (M=56.12, SD=8.10) compared to two weeks prior, before attending the workshop (M=52.62, SD=9.80); t(25)=3.48, $p$=0.002). As hypothesized, after attending the workshop the mean KAT score for participating teachers was significantly higher (ES=.68)[7]. It is also noteworthy that the standard deviation for the MSP-PD teachers dropped from 9.8 on the pre-test to 8.1 on the post-test.

One threat to the inference that the observed gains are generalizable to other mathematics teachers in the district is the potential that the self-selection of teachers into the MSP-PD resulted in a sample of teachers that were different from other teachers in the district. In particular, they could have begun the study with higher content knowledge and thus may have had greater potential to improve their knowledge, or vice-versa, they may have had lower content knowledge at the beginning providing greater room for improvement. Thus, in order to further support my hypothesis that gain in teachers' content knowledge was related to participation in the MSP-PD, I also compared MSP-PD teachers' pre-test KAT scores with the comparison group teachers' KAT scores. In fact, an independent-samples t-test indicated that there was no significant mean difference between MSP-PD teachers' pre-test score (M=52.44, SD=10.52) and comparison teachers' score (M=52.83, SD=10.01); t(42)=-0.118, $p$=0.885. This suggests that, as hypothesized, before attending the PD, MSP-PD teachers' mathematical content knowledge of algebra1 was roughly representative of a sample of comparison group teachers in the same district.

---

[7] According to the Cohen's convention, this is a large effect size. However, many researchers regard effect sizes in within-subjects designs (e.g. pre-post design) as an overestimation of the "true" effect size (e.g., Dunlap et al., 1996; Maxwell & Delaney, 2004; Olejnik & Algina, 2003, as cited in Lakens, 2013).

Although teachers were no different from a group of comparison teachers on their pre-test KAT scores, after attending the MSP-PD summer workshop teachers' mathematical content knowledge significantly improved, as hypothesized. Moreover, the standard deviation for the group also declined indicating the nature of the improvement – teachers at the lower end of the spectrum made greater gains, reducing the variability for the overall mean.

## 5.2 EFFECT OF MSP-PD ON STUDENTS' ACHIEVEMENT

I used repeated measures HLM models to examine the effect of the MSP-PD on changes in students' mathematics achievement. I ran six separate models using data from all tier1 and tier2 courses for all students in the district for the topics algebra1, geometry, and algebra2. I hypothesized that students in classrooms with a teacher who participated in the MSP-PD would demonstrate higher achievement outcomes relative to students in classroom not taught by an MSP-PD participant. By running separate models for each course, I also explored whether the MSP-PD effect varies across topics. My hypothesis was that MSP-PD would be more effective for promoting changes in students' achievement in algebra1 and geometry because those topics were more aligned with the focus of the MSP-PD. Moreover, by comparing the results from tier1 and tier2 courses, I also explored whether the MSP-PD effect was greater for low performing students, who generally comprise the tier2 courses.

When models were examined within topics at the tier level, standard errors of the MSP-PD estimates were quite large. One of the reasons for that was the small number of classrooms resulting in limited power to detect all but large effects in these models. Thus, in order to

increase the sample size and lower the standard errors, I also ran more general models[8] by combining data from tier1 and tier2 students for each topic. These general models had more power and hence they could detect smaller effects of the MSP-PD on student growth.

Results presented below were from models including strata covariates indicating matched classrooms and adjusting for student level exogenous variables including students' SES, gender and race, effectively accounting for any student differences between students within classrooms. In the tables that follow, I present only the results for coefficients for MSP-PD effects because they are of primary interest. Tables including full results of all covariates are provided for algebra1 in Appendix C, for geometry in Appendix D, and for algebra2 in Appendix E.

### 5.2.1 The effect of MSP-PD on students' algebra1 achievement

#### 5.2.1.1 Tier1 and tier2 findings

Results from the HLM growth models revealed a marginally significant MSP-PD effect on the growth rate of tier2 students in algebra1 ($\gamma_{101} = 0.20$, *SE*=0.11, p=0.097). This means that, the average change rate for MSP-PD students' scores were 0.20 standard deviations higher than comparison students' scores. The rate of .20 standard deviations corresponds to each one unit increase in time, where time increased by 1 for every CBA administration. Therefore, we

[8] Structural forms of general models are the same as structural forms of the specific models that I ran for each tier. The only difference is that, in models within tiers, there is only one stratum in the models whereas in general models there are multiple strata. I used the same two strata that I created for tier1 and tier2 models. Combining the two datasets resulted in four dummy strata variables; one stratum as a reference and three strata in the models.

observed in total a 0.60 standard deviations difference in students' rates of change in their algebra1 scores, favoring MSP-PD students during AY 2011-2012[9].

Similarly, tier1 students whose teachers attended MSP-PD also had a higher rate of change in algebra1 ($\gamma_{101} = 0.12$, *SE*=0.08, *p*=0.177) relative to comparison students, but here the difference in rates of change was not statistically significant. In terms of the random effects for the change rates in models for tier2 and tier1 students, as a predictor MSP-PD by itself explained 24% of the variation between tier1 classrooms and 20% of the variation between tier2 classrooms in these models (See Table 10).

[9] Estimated growth rate is an average increase in students' scores at each administration of tests. Using this estimation, we can calculate growth rate for the entire school year, because there are three different intervals (including the interval from September to the first CBA administration). Alternatively, by re-coding the time variable we can directly get this estimate from the model. Since both strategies generated similar growth rates for the school year, I provided results in the present metric.

**Table 10.** Effects of MSP-PD on students' algebra1 achievement in tier1 and tier2 courses

| | Algebra1 Tier1 | | | Algebra1 Tier2 | | |
|---|---|---|---|---|---|---|
| **Fixed Effects** | | | | | | |
| | Coeff. | se | p-value | Coeff. | se | p-value |
| Mean final status ($\gamma_{000}$) | -0.36 | 0.25 | 0.182 | -0.21 | 0.26 | 0.426 |
| MSP-PD Effect ($\gamma_{001}$) | 0.41 | 0.29 | 0.187 | 0.28 | 0.27 | 0.309 |
| | | | | | | |
| Mean growth rate ($\gamma_{100}$) | -0.11 | 0.07 | 0.175 | -0.12 | 0.15 | 0.431 |
| MSP-PD effect ($\gamma_{101}$) | 0.12 | 0.08 | 0.177 | 0.20 | 0.11 | 0.097 |
| **Random Effects** | | | | | | |
| Between classrooms | Variance Component | | p-value | Variance Component | | p-value |
| Final status ($u_{ooj}$) | 0.20 | | < 0.001 | 0.21 | | < 0.001 |
| Growth rate ($u_{1oj}$) | 0.01 | | 0.006 | 0.03 | | < 0.001 |
| | Variance Explained by MSP-PD | | | Variance Explained by MSP-PD | | |
| Final status | 0.15 | | | 0.05 | | |
| Growth rate | 0.24 | | | 0.20 | | |

Moreover, the MSP-PD effect on students' final achievement status in algebra1 was positive but not significant for both tier1 ($\gamma_{001} = 0.41$, *SE*=0.29, *p*=0.187) and tier2 courses ($\gamma_{001} = 0.28$, *SE*=0.27, *p*=0.309). For tier1, MSP-PD explained 15% of the between classroom variation in final achievement status in algebra1, while it only explained 5% of between classroom variation in the tier2 model (See Table 10). For tier1 and tier2 courses, students' achievement trajectories for MSP-PD and comparison students for algebra1 are depicted in Figure 1.

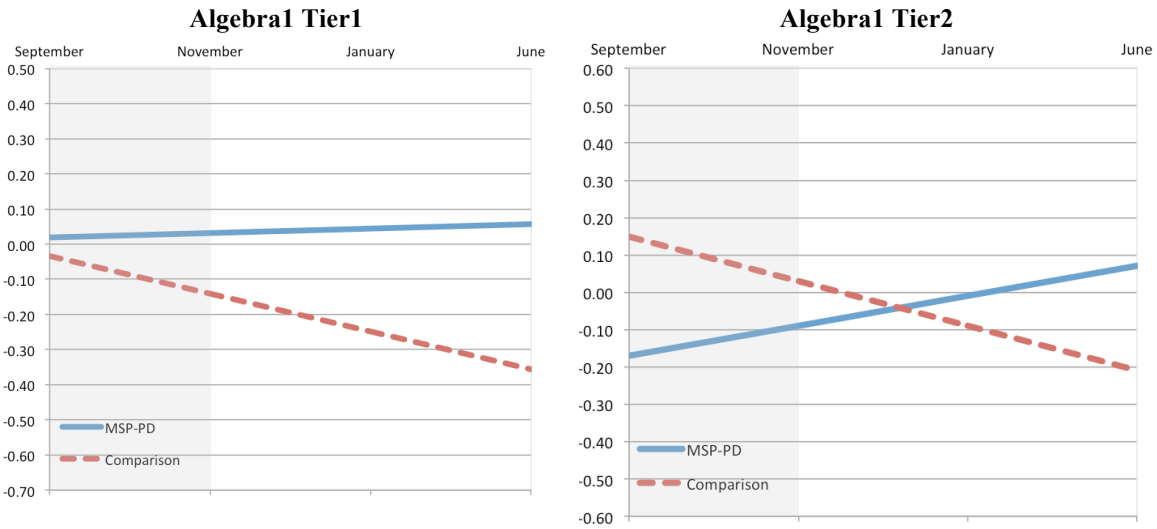



**Figure 1.** Achievement trajectories for MSP-PD and comparison group students' algebra1 scores in tier1 and tier2 courses.

### 5.2.1.2 Findings for general model (tier1 and tier2 combined):

Results of the general model for algebra1 revealed that MSP-PD students' mean rate of change in achievement over the year ($\gamma_{101} = 0.16$, *SE*=0.06) and mean final status ($\gamma_{001} = 0.34$, *SE*=0.18) were higher than comparison students. The difference between MSP-PD and comparison students was statistically significant on the rate of change ($p$=0.038) and marginally significant on final status for algebra1 ($p$=0.080). In this model, the addition of MSP-PD to the models explained 17% of the variance in the rate of change for students' achievement between all algebra1 classrooms and 9% the variance between classrooms in status (See Table 11). Trajectories for rates of change for MSP-PD and comparison students for all algebra1 courses combined are shown in Figure 2.

**Table 11.** Effects of MSP-PD on students' algebra1 achievement across tier1 and tier2 courses (combined model)

| **Fixed Effects** | | | |
|---|---|---|---|
| | Coeff. | se | p-value |
| Mean final status ($\gamma_{000}$) | -0.52 | 0.22 | 0.023 |
| MSP-PD Effect ($\gamma_{001}$) | 0.34 | 0.18 | 0.080 |
| | | | |
| Mean growth rate ($\gamma_{100}$) | -0.15 | 0.14 | 0.202 |
| MSP-PD effect ($\gamma_{101}$) | 0.16 | 0.06 | 0.038 |
| **Random Effects** | | | |
| Between classrooms | Variance Component | | p-value |
| Final status ($u_{ooj}$) | 0.20 | | < 0.001 |
| Growth rate ($u_{1oj}$) | 0.03 | | < 0.001 |
| | Variance Explained by MSP-PD | | |
| Final status | 0.09 | | |
| Growth rate | 0.17 | | |

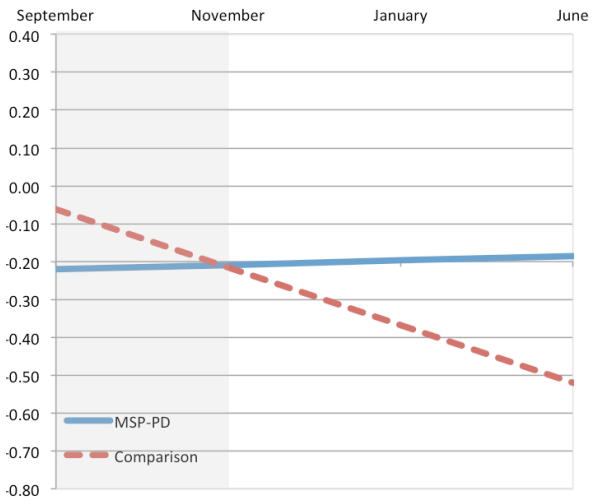



**Figure 2.** Achievement trajectories for MSP-PD and comparison group students' algebra-1 scores across tier1 and tier2 courses.

Summary:

Findings revealed a marginally significant effect of MSP-PD on student's rates of change for their achievement in algebra1 when tier2 classrooms were examined by themselves. Moreover, the general model for algebra1 revealed that MSP-PD had a significant effect on students' rates of change for their achievement and a marginally significant effect on students' final achievement status when tier1 and tier2 courses were combined.

### 5.2.2 The effect of MSP-PD on students' geometry achievement

#### 5.2.2.1 Tier1 and tier2 findings

The MSP-PD had a significant effect on changes in achievement trajectories of tier2 students ($\gamma_{101} = 0.27$, $SE$=0.09, $p$=0.047) and nearly a marginally significant effect on their final achievement status in geometry ($\gamma_{001} = 0.58$, $SE$=0.28, $p$=0.104). This suggest that, MSP-PD

students' mean score of geometry changed 0.81 standard deviation more relative to comparison students' scores. This difference in rates of change for achievement trajectories resulted in a 0.58 standard deviation difference for MSP-PD students relative to comparison students at the end of the school year in the geometry tier2 course. Random effects from these models also revealed a substantial effect of MSP-PD. 47% of the between classroom variation at final status and 84% of the between classroom variation in changes in achievement trajectories were explained by MSP-PD only in this model (See Table 12).

**Table 12.** Effects of MSP-PD on students' geometry achievement in tier1 and tier2 courses

| | Geometry Tier1 | | | Geometry Tier2 | | |
|---|---|---|---|---|---|---|
| **Fixed Effects** | | | | | | |
| | Coeff. | se | p-value | Coeff. | se | p-value |
| Mean final status ($\gamma_{000}$) | -0.04 | 0.21 | 0.853 | 0.28 | 0.34 | 0.458 |
| MSP-PD Effect ($\gamma_{001}$) | 0.18 | 0.23 | 0.444 | 0.58 | 0.28 | 0.104 |
| | | | | | | |
| Mean growth rate ($\gamma_{100}$) | 0.00 | 0.11 | 0.988 | -0.14 | 0.19 | 0.500 |
| MSP-PD effect ($\gamma_{101}$) | 0.01 | 0.11 | 0.935 | 0.27 | 0.09 | 0.047 |
| **Random Effects** | | | | | | |
| Between classrooms | Variance Component | | p-value | Variance Component | | p-value |
| Final status ($u_{ooj}$) | 0.15 | | < 0.001 | 0.11 | | < 0.001 |
| Growth rate ($u_{1oj}$) | 0.03 | | < 0.001 | 0.00 | | 0.036 |
| | Variance Explained by MSP-PD | | | Variance Explained by MSP-PD | | |
| Final status | 0.05 | | | 0.47 | | |
| Growth rate | 0.00 | | | 0.84 | | |

In contrast to tier2, findings for the geometry tier1 course revealed that the MSP-PD had no effect on either changes in students' achievement trajectories ($\gamma_{101} = 0.01$, *SE*=0.11, $p$=0.935) or on students' final achievement status ($\gamma_{001} = 0.18$, *SE*=0.23, $p$=0.444). For the tier1 model, MSP-PD explained no between classrooms variation in students' rates of change and only 5% of the between classrooms variation in final status (See Table 12). Growth trajectories are depicted for tier1 and tier2 courses in Figure 3.

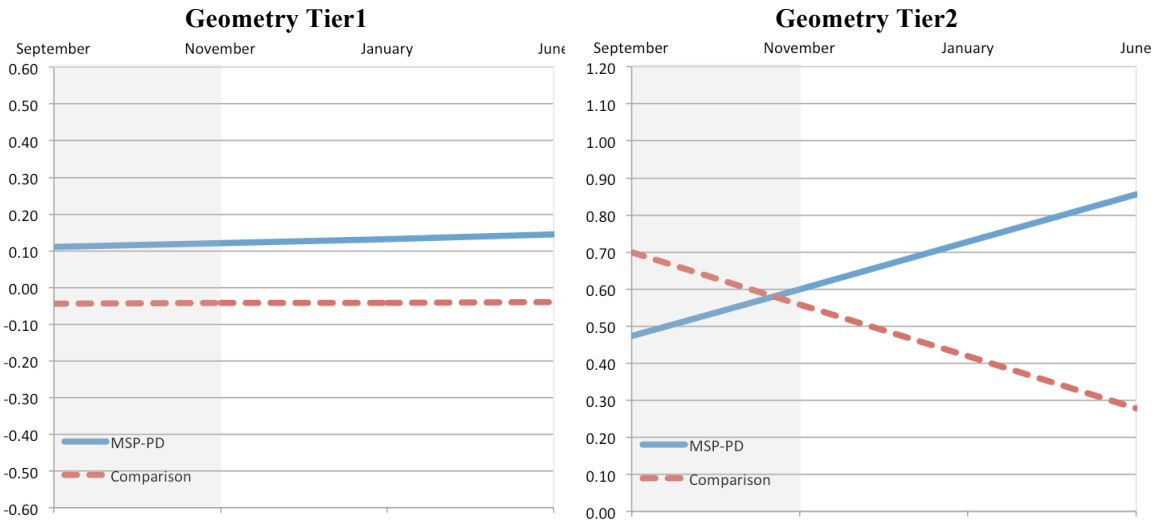



**Figure 3.** Achievement trajectories for MSP-PD and comparison group students' geometry scores in tier1 and tier2 courses.

### 5.2.2.2 Findings for general model (tier1 and tier2 combined):

Lastly, results of the general model for geometry revealed that MSP-PD students' mean rates of change ($\gamma_{101} = 0.11$, *SE*=0.06, $p$=0.168) and mean final status ($\gamma_{001} = 0.22$, *SE*=0.17, $p$=0.211) were higher than comparison students. However, these differences were not statistically significant. In this model, MSP-PD by itself explained 13% of the variance between classrooms

in growth and 10% in final status of all students in geometry (See Table 13). Achievement trajectories for geometry are depicted in Figure 4.

**Table 13.** Effects of MSP-PD on students' geometry achievement across tier1 and tier2 courses (combined model)

| **Fixed Effects** | | | |
|---|---|---|---|
| | Coeff. | se | p-value |
| Mean final status ($\gamma_{000}$) | -0.18 | 0.25 | 0.480 |
| MSP-PD Effect ($\gamma_{001}$) | 0.22 | 0.17 | 0.211 |
| | | | |
| Mean growth rate ($\gamma_{100}$) | -0.02 | 0.13 | 0.901 |
| MSP-PD effect ($\gamma_{101}$) | 0.11 | 0.08 | 0.168 |
| **Random Effects** | | | |
| Between classrooms | Variance Component | | p-value |
| Final status ($u_{ooj}$) | 0.12 | | < 0.001 |
| Growth rate ($u_{1oj}$) | 0.03 | | < 0.001 |
| | Variance Explained by MSP-PD | | |
| Final status | 0.10 | | |
| Growth rate | 0.13 | | |

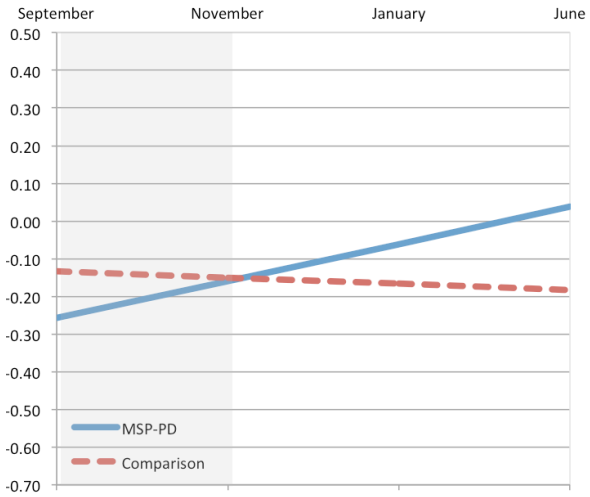



**Figure 4.** Achievement trajectories for MSP-PD and comparison group students' geometry scores across tier1 and tier2 courses.

Summary:

To sum up, for tier2 students, MSP-PD had a significant effect on the average rate of change of students' achievement trajectories. As the magnitude of the estimated coefficient for MSP-PD indicates (ES=.81) and the reduction in variance between classrooms revealed, MSP-PD had a large effect on rates of change of tier2 students' achievement in geometry during AY 2011-2012. However, in the general model, when data from tier1 and tier2 courses were combined, there was no significant effect of MSP-PD on student achievement.

### 5.2.3 The effect of MSP-PD on students' algebra2 achievement

#### 5.2.3.1 Tier1 and tier2 findings

Results from the repeated measures models found no effect of MSP-PD on students' rates of change in algebra2 for both tier1 ($\gamma_{101} = 0.00$, $SE$=0.06, $p$=0.984) and tier2 courses ($\gamma_{101} = -0.12$, $SE$=0.21, $p$=0.574). Similarly, MSP-PD had no effect on final achievement status of algebra2 for

tier1 ($\gamma_{001}$ = -0.01, *SE*=0.16, *p*=0.939) and tier2 courses ($\gamma_{001}$ = -0.17, *SE*=0.36, *p*=0.647). Moreover, almost no variation was explained by MSP-PD in these models (See Table 14). Changes in rates of students' achievement trajectories for MSP-PD and comparison group students' algebra2 scores in tier1 and tier2 courses are shown in Figure 5.

**Table 14.** Effects of MSP-PD on students' algebra2 achievement in tier1 and tier2 courses

| | Algebra2 Tier1 | | | Algebra2 Tier2 | | |
|---|---|---|---|---|---|---|
| **Fixed Effects** | | | | | | |
| | Coeff. | se | p-value | Coeff. | se | p-value |
| Mean final status ($\gamma_{000}$) | -0.23 | 0.13 | 0.098 | -0.09 | 0.21 | 0.688 |
| MSP-PD Effect ($\gamma_{001}$) | -0.01 | 0.16 | 0.939 | -0.17 | 0.36 | 0.647 |
| | | | | | | |
| Mean growth rate ($\gamma_{100}$) | 0.01 | 0.04 | 0.802 | 0.01 | 0.12 | 0.913 |
| MSP-PD effect ($\gamma_{101}$) | 0.00 | 0.06 | 0.984 | -0.12 | 0.21 | 0.574 |
| **Random Effects** | | | | | | |
| Between classrooms | Variance Component | | p-value | Variance Component | | p-value |
| Final status ($u_{ooj}$) | 0.08 | | < 0.001 | 0.21 | | < 0.001 |
| Growth rate ($u_{1oj}$) | 0.00 | | 0.013 | 0.07 | | < 0.001 |
| | Variance Explained by MSP-PD | | | Variance Explained by MSP-PD | | |
| Final status | 0.00 | | | 0.00 | | |
| Growth rate | 0.00 | | | 0.02 | | |

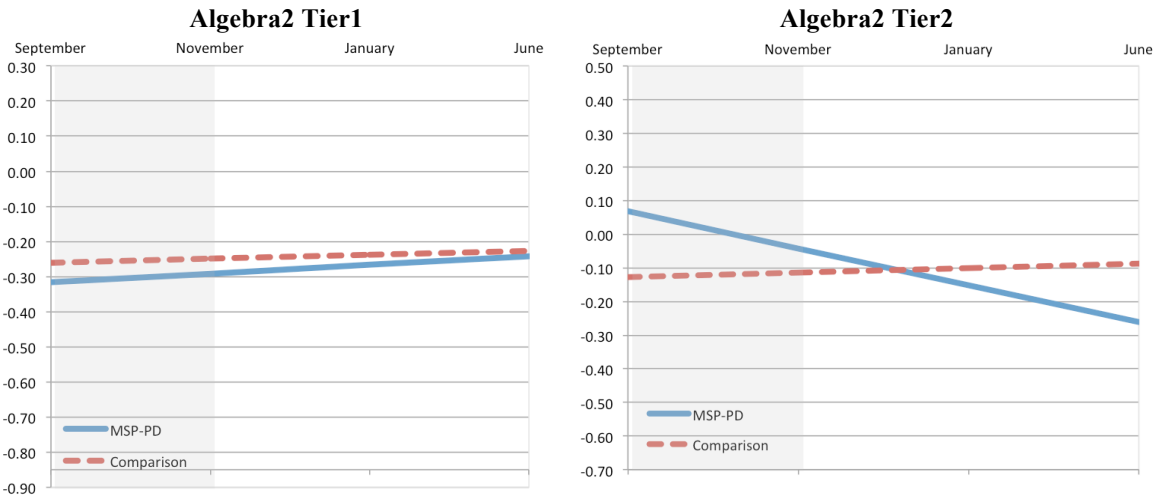



**Figure 5.** Achievement trajectories for MSP-PD and comparison group students' algebra2 scores in tier1 and tier2 courses.

### 5.2.3.2 Findings for general model (tier1 and tier2 combined):

In the combined model, MSP-PD students' mean rates of change for achievement trajectories ($\gamma_{101}$ = -0.04, *SE*=0.08, *p*=0.662) and mean final status ($\gamma_{001}$ = -0.05, *SE*=0.16, *p*=0.770) were not significantly different from comparison students (See Table 15). In this model, MSP-PD explained only 2% of the variance between classrooms in growth and 1% in final status of all students in algebra2. Achievement trajectories for MSP-PD and comparison students for algebra-2 across tier1 and tier2 courses are shown in Figure 6. These results suggest that MSP-PD had no effect on students' algebra2 achievement.

**Table 15.** Effects of MSP-PD on students' algebra2 achievement across tier1 and tier2 courses (combined model)

| **Fixed Effects** | | | |
|---|---|---|---|
| | Coeff. | se | p-value |
| Mean final status ($\gamma_{000}$) | -0.07 | 0.14 | 0.631 |
| MSP-PD Effect ($\gamma_{001}$) | -0.05 | 0.16 | 0.770 |
| | | | |
| Mean growth rate ($\gamma_{100}$) | 0.04 | 0.07 | 0.589 |
| MSP-PD effect ($\gamma_{101}$) | -0.04 | 0.08 | 0.662 |
| **Random Effects** | | | |
| Between classrooms | Variance Component | | p-value |
| Final status ($u_{ooj}$) | 0.12 | | < 0.001 |
| Growth rate ($u_{1oj}$) | 0.03 | | < 0.001 |
| | Variance Explained by MSP-PD | | |
| Final status | 0.01 | | |
| Growth rate | 0.02 | | |

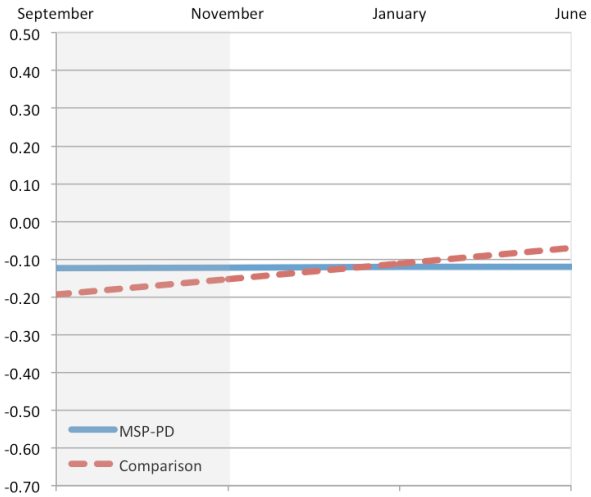



**Figure 6.** Achievement trajectories for MSP-PD and comparison group students' algebra2 scores across tier1 and tier2 courses.

# 6.0 DISCUSSION

Results of this quasi-experimental study indicated that the MSP-PD was effective for improving student achievement in algebra1 (across tier1 and tier2 courses) and it was also effective for improving student achievement in the tier 2 geometry course. MSP-PD influenced not only the final status of student achievement but the changes in achievement trajectories, as well. This means that, within a school year, performances of students in MSP-PD classrooms were improving in CBAs assessments relative to the performances of students in comparison classrooms. Moreover, the differences in achievement trajectories of MSP-PD and comparison students were substantial in these courses. Extrapolating "growth" rate to an academic year revealed a medium-sized MSP-PD effect in algebra1 across tier1 and tier2 courses ($d$= 0.48 in combined model) and a large effect in geometry tier2 courses ($d$= 0.81). For the effect of MSP-PD on students' final achievement status, there was a small-medium effect in algebra1 ($d$= 0.34) and medium-high effect in the geometry tier2 courses ($d$= 0.58).

Compared to detected effect sizes of the PD programs in prior experimental/quasi-experimental studies, the effect sizes of MSP-PD on student achievement observed here, especially on the changes in students' achievement trajectories, were quite large in this study. Average effect sizes of the PD programs in review studies of prior experimental/quasi-experimental studies ranged from small (d= 0.14) to medium (d=0.54) (Ball et al., 2008; Blank &

de las Alas, 2009; Clewell et al., 2004; Kennedy, 1998; Scher & O'Reilly, 2009; Yoon et al., 2007).

## 6.1 IMPLICATIONS: DESIGNING EFFECTIVE PD PROGRAMS AND STUDIES WITH EFFECTIVE RESEARCH DESIGNS

By demonstrating the effectiveness of the MSP-PD, this study attempted to address the shortage of robust empirical evidence for the effectiveness of PD on student achievement. Relative to the large number of PD studies over the past 15 years, the number of experimental/quasi-experimental studies of PD effects on achievement is small. Moreover, because it is very difficult to isolate PD effects from other external factors potentially confounding the effects of PD, very few of these studies could demonstrate the effect of PD on student achievement. As a result, the knowledge base for features of effective PD programs is mostly based on theory, correlational survey studies, and case studies (AERA, 2005; Scher & O'Reilly, 2009).

By demonstrating an effect of the MSP-PD on student achievement, this study contributes to our research-based knowledge about some features of effective PD programs. Moreover, it provides a case study showing how the research design might contribute in important ways to the ability to detect an effect of PD on student achievement. Thus, this study has implications for multiple audiences. For one, the findings speak to both educational practitioners and policymakers in their efforts to design and support effective PD programs. Furthermore, for educational researchers the findings provide evidence for potential strategies for demonstrating robust research-based evidence for the effectiveness of PD on student learning.

### 6.1.1 PD programs should have effective design features in order to improve student achievement.

One of the central findings of this study is the importance of the PD having a content focus (in this case mathematics) in order to improve student achievement outcomes. This finding informs both educational practitioners as well as educational researchers. The MSP-PD focused on teachers' actively engaging with each other and with course facilitators in doing mathematics in the topic areas algebra1 and geometry. Teachers were both working on content as mathematicians and engaging in mathematics as a discipline. They reflected on their own disciplinary experiences as learners and discussed implications of creating similar learning opportunities for their students. Thus, MSP-PD was also focused on specialized content knowledge of how to teach the content effectively.

In addition, the content of the MSP-PD was closely linked with the curriculum. Review studies have shown that PD programs that were tied to curriculum, to knowledge of subject matter, and/or to how students learn the subject have been more effective for improving student achievement and have produced, on average, larger effect sizes (Clewell et al., 2004; Kennedy 1998). The findings shown here for the effect of the MSP-PD on student achievement provide further evidence to support this already established line of evidence. As discussed in the literature review, PD containing a content focus is the only feature of PD that prior experimental/quasi-experimental studies have provided strong empirical evidence for.

Another design feature of the MSP-PD was that it was intense in its duration, was coherent and extended over a full year[10]. It included a 2-week summer institute and 6 follow-up

[10] With several teachers participating in more than one consecutive year(s).

sessions during the school year totaling 110 hours of PD to teachers. Although the empirical evidence for the effectiveness of duration and high contact hours of PD has provided mixed evidence in the literature, most of the effective PD programs examined in recent experimental/quasi-experimental studies all consisted of a longer duration and high contact hours. The MSP-PD studied here also fit this profile because it was effective in improving student achievement and it had long duration and high contact hours. Thus, the findings of this study also contribute to the growing body of evidence that PD of intense duration, coherent and provided for an extended time period is more likely to have an effect on student achievement.

This finding suggests that, it is important to ensure that the PD program is content-focused, intense, and provided over longer time period while designing a PD program or deciding which PD interventions to support. This has implications for educational practitioners and policymakers. Unfortunately, as national survey studies have continually revealed, large numbers of teachers are still attending brief, one-shot workshops despite consensus among scholars that such PD programs are not effective for improving student achievement (Birman et al., 2001; Darling-Hammond et al., 2009; Hill, 2007; Parsad et al., 2001) and growing empirical evidence that content-focused intensive PD is effective. Thus, more needs to be done to educate educational practitioners and policymakers about the importance of some central features of PD efforts. Further, resources should be provided to guide practitioners and policymakers toward more effective PD designs. For educational researchers, meanwhile, the findings of this study underscore the importance of ensuring that the PD programs under examination have central design features before examining its effect. If the PD fails to contain these features, assessing its effectiveness with well-designed research likely will not help to find an effect of the PD.

### 6.1.2 Focus of the PD and its alignment with the outcome measure matters for PD effectiveness and PD research.

In this study, I hypothesized that the alignment of the topic coverage of the PD would result in different effect sizes for the different topics examined. Specifically, intense and ongoing PD focusing on algebra1 and geometry would improve teachers' instruction especially for those subjects. Consequently, improvements in students' achievement in algebra1 and geometry would be more easily detectable. As hypothesized, this study found an effect of the MSP-PD on changes in students' achievement trajectories in algebra1 and geometry tier2 courses, but teachers' participation in the MSP-PD did not influence changes in students' achievement trajectories in algebra2.

This finding might suggest that the more PD is narrowed to specific content areas and is closely linked to the curriculum, the more it is effective for improving teaching and student achievement in that targeted content. However, it might also indicate that it is difficult for teachers to transfer the acquired knowledge and skills to another content area[11].

Regardless of underlying factors leading to the null findings for the effect of MSP-PD on algebra2, this finding emphasizes the importance of using outcome measures aligned with and sensitive to the focus of the PD. This has important implications for educational researchers. Review studies have indicated that PD studies utilizing student outcome measures aligned with the focus of the PD found larger effect sizes of the PD on student achievement. Thus, using CBAs that are sensitive to and closely aligned with the focus of the PD might help to better

[11] This would need to be examined further especially because the sample size investigated in this one district prohibits making any conclusions about the extent to which PD in one topic area transfers to teacher learning and proclivity to improve teaching in other content areas.

detect PD effects on student achievement and to detect larger effect sizes for PD. Moreover, using CBAs resulted in a coherent system where the PD, curriculum, and assessments were all aligned. That also could have significantly contributed to properly capturing the MSP-PD effect and has further implications for the design of effective PD as well as the design of effective research programs examining PD effects on achievement.

### 6.1.3 Demonstrating an effect of high quality PD on student achievement requires carefully designed research

This study also demonstrated that once PD is designed with effective features, detecting effects of the PD on student achievement is possible through carefully designed research. One way this is possible is by using student outcome measures aligned with the focus of the study, as discussed above. Another way is by isolating the PD effect from confounding variables. It was achieved in this study by running specific models for each course and matching classrooms within the same courses through rigorous quasi-experimental methods of propensity score stratification. As a result, at the unit of analysis-classrooms-the only differences between classrooms remained whether the teacher received treatment or not. In addition to helping to detect an effect of the MSP-PD on student achievement, this design feature also allowed me to be more confident in attributing the observed effect on student achievement to the MSP-PD (and not to other potentially confounding factors such as curriculum or student and aggregate classroom ability both of which could easily influence teaching). Ruling out alternative explanations for improvement in student achievement is challenging in PD research, but as this finding suggests it can help to generate robust evidence for the effectiveness of PD. This has implications for educational researchers in their efforts to address the shortage of robust

empirical evidence for the effectiveness of PD and when looking across studies to find patterns for which PD features seem especially important contributors for improving student achievement.

Another effective design feature of this study, which has implications for educational researchers, is using a repeated measures design to examine the effect of PD on student achievement. By using longitudinal models the effect of MSP-PD on students' achievement trajectories were estimated separately from their final achievement status. Moreover, in these models both the nested structure of the data and differences in student characteristics within classrooms were accounted for. As the literature review demonstrated, recent experimental/quasi experimental studies have moved beyond cross-sectional designs but the majority of them have used covariate-adjusted as opposed to repeated measures models. In covariate-adjusted models it is difficult to separate student's status from their growth (Rowan et al., 2002). Thus, to have a better understanding of how PD influences changes in student achievement over time, it is important for educational researchers to collect longitudinal data and employ repeated measures models to examine the effectiveness of PD programs. This study benefitted from a unique data set because (apart from the research) the district had previously decided to employ a sequence of CBAs aligned with their curriculum in the context of instituting reforms to improve their mathematics teaching.

### 6.1.4 Designing effective PD research involves considering limitations and tradeoffs associated with design features.

As I discussed previously in the literature review, it is hard for PD studies to produce an optimal research design. Design features of PD studies involve trade-offs and there are always limitations

as a result of a shortage of resources available to researchers and/or limitations associated with the complex nature of educational settings.

Some of the key design features of this study also had some accompanying drawbacks. For example, running growth models specific to each course had a substantial role in matching and evaluating the MSP-PD effect in similar classrooms. But, on the other hand, it also resulted in a small number of level-2 units in the growth models. As a result, standard errors of the MSP-PD estimate were quite large. Correspondingly, the minimum detectable effect size for the treatment on growth in achievement was also quite large[12] and it became harder to detect an effect in hypothesis testing.

By running growth models specific to each course, I could also compare and contrast the extent to which the MSP-PD influenced the achievement of students who vary in their prior mathematics achievement. Results revealed that, in algebra1 courses, both groups of students benefitted from the MSP-PD. But, for geometry, MSP-PD significantly improved the performance of low achieving students while it had no effect for high achieving students. During the academic year of this study, a new geometry curriculum had been introduced in the district. Thus, the MSP-PD effect, together with the new curriculum, might have helped low performing students accelerate even more under these conditions, However since there is limited data, I couldn't further explore why this is the case.

Using the CBAs as an outcome had several benefits for this research but there were also drawbacks associated with it. One of these drawbacks was that large variation existed in students' achievement between each administration of the CBAs. It suggests that CBAs likely do not have an optimal means for reliably measuring changes in achievement while minimizing

[12] The minimum detectable effect size (d) for algebra1 tier1 was 0.5, d=0.4 for Algebra1 tier2, d=0.5 for Geometry tier1, and d= 0.8 for Geometry tier2.

measurement error. Combined with small sample sizes at the classroom level in the growth models, these large within student variations made the minimum detectable effect size for finding effects of MSP-PD on student achievement larger. Moreover, because the CBAs were measuring students' achievement in specific content areas, it is difficult to generalize the MSP-PD effects beyond students' algebra1 and geometry achievement.

By comparison, alternative assessments would have been even more problematic. For example, if I had chosen to examine the effects by using the state test (a summative test administered in March) it might have been possible to generalize the effect beyond algebra1 and geometry but the state test data was available only for $4^{th}$ through $8^{th}$ grade students and $11^{th}$ graders, and therefore, a relatively small proportion of students in the focal topics. Moreover, using such a summative assessment would require examining the MSP-PD effect solely through a cross-sectional analysis and only at the given time point. Additionally, if I employed covariate-adjusted models using state tests, it would rely on gains in student achievement between March 2011 and March 2012 (including two months of teaching of different teacher and leaving two months of teaching of the current teacher) thus would result in less precise estimation of an MSP-PD effect.

Another design feature that had demonstrated benefits but also limitations was the examination of the MSP-PD effect on my proximal teacher outcome, that is, teachers' content knowledge. While I couldn't examine the effect of MSP-PD on teachers' instruction since this data was not available, showing the improvement in teachers' content knowledge after they completed the MSP-PD workshop helped partially confirm the theory of action for the MSP-PD. Hence, it helped to increase the confidence in attributing the observed effect on student achievement to the MSP-PD. However, the improvements in teacher knowledge were related

solely to the effect of the MSP-PD workshop because the post-test was given prior to the follow-up sessions. Possible gains in MSP-PD teachers' content knowledge that occurred later in the school year as a result of further learning during the follow-up sessions were not captured in this data. Moreover, the statistical analyses used to examine the MSP-PD effect on teacher knowledge were not as rigorous as the achievement analyses. I employed both cross-sectional and a pre-post analysis of group means, but no control variables were used in these analyses because information about the teachers and their contextual factors was not available. Furthermore, although comparison group teachers were recruited, they were only administered the knowledge test once in the late fall. Thus, I couldn't directly compare MSP-PD teachers' gains against the comparison group teachers' gains because no gain score could be calculated for the comparison group. Moreover, comparison group teachers for this analysis volunteered to take knowledge test. Like MSP-PD teachers, comparison group teachers might also be different from other teachers in the district.

Lastly, standardizing students' scores at each time point and using these standardized scores as an outcome, instead of using raw scores, had clear benefits for examining the effect of MSP-PD on student achievement trajectories. Using standardized scores at each administration of the test adjusts for the test difficulty and equates scores for each CBA administration. When standardized scores are used, the growth model examines how MSP-PD students' standing changes relative to students in the comparison group over time. However, if item level data were available, I could have run measurement models in order to understand whether students' ability improved over time rather than simply make inferences about their relative performance. This would have provided even more reliable and relevant estimation of the MSP-PD effect.

Summary:

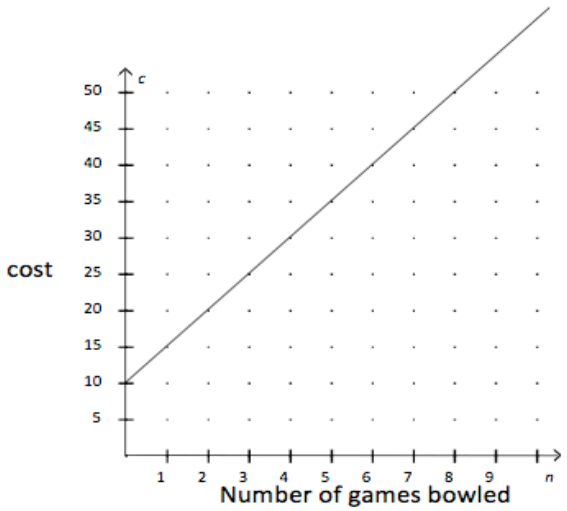
All in all, by using resources available to the researcher this study was able to demonstrate an effect for content focused, intense, ongoing PD. The research design allowed me to isolate the effect of the MSP-PD from other confounding factors such as the curriculum and students' ability. As a result, the findings provide strong empirical evidence indicating that content focused, intense, ongoing mathematics PD was effective for improving students' achievement trajectories (versus a matched comparison sample) aligned with the content focus of the PD.

# APPENDIX A

# SAMPLE CBA ITEMS

## A.1 A SAMPLE ITEM FROM THE FIRST CBA ALGEBRA1 TIER1 TEST ADMINISTERED ON NOVEMBER IN 2011-2012 SCHOOL YEAR.

Andrea belongs to a bowling league. The graph below shows the total amount she has paid, based on the number of games she bowls.




Which equation for the cost (c) based on the number of games (n) represents this situation?

A. $c = 5n$

B. $c = n + 5$

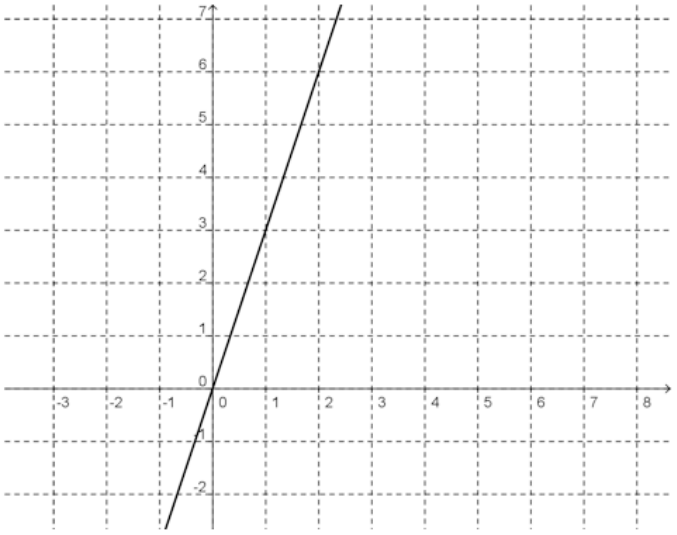
C. $c = 10n + 5$

D. $c = 5n + 10$

**Content standard:** Write and/or solve a system of linear equations (including problem situations) using graphing, substitution, and/or elimination.

## A.2 A SAMPLE ITEM FROM THE SECOND CBA ALGEBRA1 TIER1 TEST ADMINISTERED ON JANUARY IN 2011-2012 SCHOOL YEAR.

Consider the graph of the line shown below:



Which equation models the relationship between x and y as shown on the graph.

A. $y = x + 3$

B. $y = 3x + 3$

C. $y = \frac{1}{3}x$

D. $y = 3x$

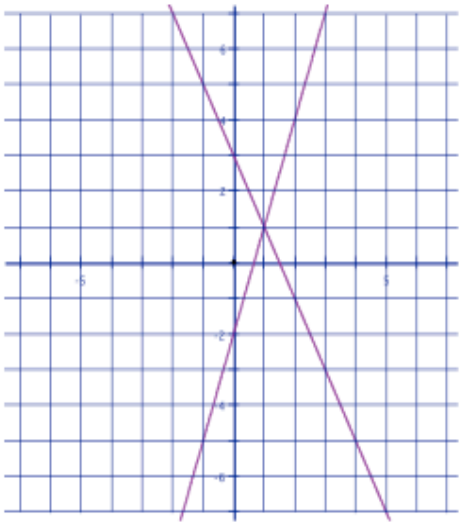
**Content standard:** Create, interpret, and/or use the equation, graph, or table of a linear function.

## A.3 A SAMPLE ITEM FROM THE THIRD CBA ALGEBRA1 TIER1 TEST ADMINISTERED ON JUNE IN 2011-2012 SCHOOL YEAR.

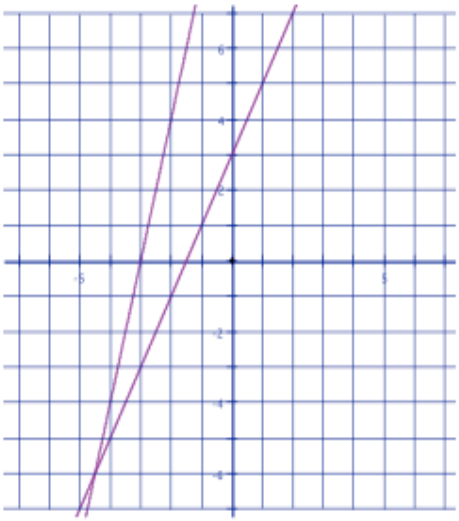
Which of the following graphs displays the solution to the system of equations below?

$$y = -2x + 3$$
$$y = 3x - 2$$

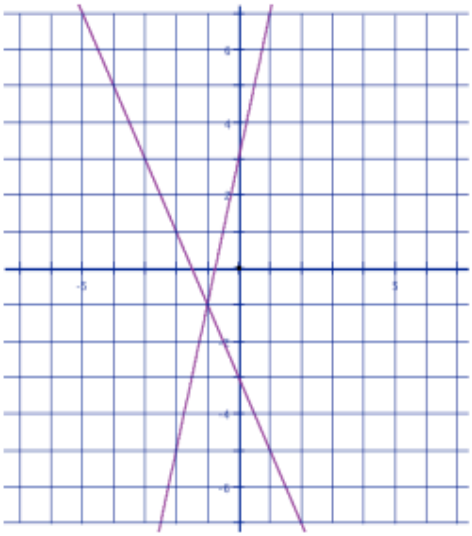
A.



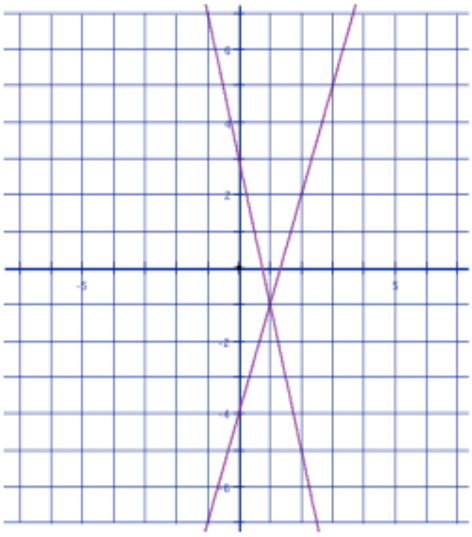
C.



B.



D.



**Content standard:** Write and/or solve a system of linear equations (including problem situations) using graphing, substitution, and/or elimination.

# APPENDIX B

## SELECTED OBSERVED COVARIATES FOR BALANCE CHECK AFTER MATCHING

| **Classroom Level Covariates** | **Teacher Level Covariates** | **School Level Covariates** |
|---|---|---|
| Percent of Male Students | Male | Number of teachers |
| *Students' Race* | White | Number of students |
| Percent of White Students | Black | Student Teacher Ratio |
| Percent of Black Students | Asian | Percent of Students Eligible for Free or Reduced Lunch |
| Percent of Hispanic Students | Other | Mean Prior Math Achievement |
| Percent of Asian Students | Age | Percent Black Students |
| Percent of Other Race Students | *Teacher's Instructional Role* | Percent White Students |
| *Students' Grade Level* | Middle School Teacher | Percent Hispanic Students |
| Percent of 12$^{th}$ Grade Students | Secondary School Teacher | Percent Asian Students |
| Percent of 11$^{th}$ Grade Students | Promise-Readiness Corps | Percent Other Students |
| Percent of 10$^{th}$ Grade Students | Instructional Teacher Leader | Percent of Stable Students |
| Percent of 9$^{th}$ Grade Students | Clinical Resident Instructors | Number of Incidents |
| Percent of 8$^{th}$ Grade Students | Substitute Teacher | Number of Expulsions |
| Mean Student Age | Regular Teacher | Mean Attendance Rate |
| Mean Prior Math Achievement | Provisional Pathway | Number of Expulsions |
| Percent of Students Eligible for Free or Reduced Lunch | Other Instructional Role | Percent of Male Teachers |
| Mean Attendance Rate | Graduate Degree | Percent of Teachers with Graduate |

| | | |
|---|---|---|
| | | Degree |
| Mean Enrollment Rate | NBPTS (professional teaching standards) Certificate | Percent of Teachers with NBPTS Certificate |
| | Years of Experience | Mean Teacher Experience |
| | | Mean Teacher Age |

# APPENDIX C

## RESULTS FROM ALL ALGEBRA1 GROWTH MODELS

## C.1 RESULTS FROM ALGEBRA1 TIER1 MODEL

**Table 1.** Analysis of tier1 students' algebra1 achievement with three level hierarchical growth model.

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.36 | | 0.25 |
| *Classroom Level* | | | |
| MSP-PD | 0.41 | | 0.29 |
| Strata1 | 0.20 | | 0.29 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.06 | | 0.12 |
| Black | 0.28 | | 0.29 |
| White | 0.54 | ~ | 0.30 |
| Hispanic | 0.01 | | 0.53 |
| Asian | 1.52 | *** | 0.40 |
| Female | 0.04 | | 0.10 |
| **Linear growth rate** | -0.11 | | 0.07 |
| *Classroom Level* | | | |
| MSP-PD | 0.12 | | 0.08 |
| Strata1 | 0.06 | | 0.08 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.08 | | 0.07 |
| White | 0.27 | | 0.17 |
| Black | 0.27 | | 0.17 |
| Hispanic | 0.10 | | 0.32 |
| Asian | 0.63 | ** | 0.24 |
| Female | -0.01 | | 0.06 |

| **Random Effects** | *Variance Component* | | *df* |
|---|---|---|---|
| *Level 1 (Time)* | | | |
| Temporal variation | 0.41 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.39 | *** | 302 |
| Growth rate | 0.03 | ** | 302 |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.20 | *** | 8 |
| Growth rate | 0.01 | ** | 8 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 2.** Variance decompositions for unconditional, adjusted, and final models for algebra1 tier1.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.41 | 0.41 | 0.41 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.46 | 0.39 | 0.39 |
| Growth rate | 0.04 | 0.03 | 0.03 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.25 | 0.23 | 0.20 |
| Growth rate | 0.01 | 0.01 | 0.01 |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.15 | 0.00 |
| Growth rate | -- | 0.19 | 0.02 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.05 | 0.15 |
| Growth rate | -- | 0.03 | 0.24 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.35 | -- | -- |
| Growth rate | 0.23 | -- | -- |

## C.2 RESULTS FROM ALGEBRA1 TIER2 MODEL

**Table 3.** Analysis of tier2 students' algebra1 achievement with three level hierarchical growth model.

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.21 | | 0.26 |
| *Classroom Level* | | | |
| MSP-PD | 0.28 | | 0.27 |
| Strata1 | -0.45 | | 0.26 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | 0.03 | | 0.10 |
| Black | 0.19 | | 0.19 |
| White | 0.43 | * | 0.21 |
| Hispanic | 0.48 | | 0.38 |
| Asian | 0.97 | | 0.36 |
| Female | 0.08 | | 0.08 |
| **Linear growth rate** | -0.12 | | 0.15 |
| *Classroom Level* | | | |
| MSP-PD | 0.20 | ~ | 0.11 |
| Strata1 | 0.02 | | 0.11 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | 0.12 | ~ | 0.07 |
| White | -0.01 | | 0.13 |
| Black | -0.01 | | 0.14 |
| Hispanic | 0.67 | * | 0.27 |
| Asian | 0.43 | ~ | 0.24 |
| Female | 0.07 | | 0.05 |

| **Random Effects** | *Variance Component* | | *df* |
|---|---|---|---|
| *Level 1 (Time)* | | | |
| Temporal variation | 0.48 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.22 | *** | 436 |
| Growth rate | -- | | -- |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.21 | *** | 14 |
| Growth rate | 0.03 | *** | 14 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 4.** Variance decompositions for unconditional, adjusted, and final models for algebra1 tier2.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.48 | 0.47 | 0.47 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.16 | 0.15 | 0.15 |
| Growth rate | 0.02 | 0.01 | 0.01 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.29 | 0.22 | 0.21 |
| Growth rate | 0.04 | 0.03 | 0.03 |
| | | | |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.06 | 0.00 |
| Growth rate | -- | 0.37 | 0.04 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.23 | 0.05 |
| Growth rate | -- | 0.21 | 0.20 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.64 | -- | -- |
| Growth rate | 0.70 | -- | -- |

## C.3 RESULTS FROM GENERAL ALGEBRA1 MODEL

**Table 5.** Analysis of students' algebra1 achievement across tier1 and tier2 courses with three level hierarchical growth model (Combined model).

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.52 | * | 0.22 |
| *Classroom Level* | | | |
| MSP-PD | 0.34 | ~ | 0.18 |
| Strata1 | -0.46 | ~ | 0.24 |
| Strata2 | 0.38 | | 0.25 |
| Strata3 | 0.15 | | 0.24 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.06 | | 0.07 |
| Black | 0.28 | ~ | 0.14 |
| White | 0.54 | ** | 0.15 |
| Hispanic | 0.01 | ~ | 0.27 |
| Asian | 1.52 | *** | 0.23 |
| Female | 0.04 | | 0.05 |
| **Linear growth rate** | -0.15 | | 0.14 |
| *Classroom Level* | | | |
| MSP-PD | 0.16 | * | 0.07 |
| Strata1 | -0.01 | | 0.10 |
| Strata2 | 0.01 | | 0.10 |
| Strata3 | -0.07 | | 0.10 |

| | | | |
|---|---|---|---|
| *Student Level* | | | |
| Eligible for free/reduced lunch | 0.01 | | 0.05 |
| White | 0.10 | | 0.10 |
| Black | 0.10 | | 0.10 |
| Hispanic | 0.37 | * | 0.19 |
| Asian | 0.39 | * | 0.15 |
| Female | 0.02 | | 0.04 |
| **Random Effects** | *Variance Component* | | *df* |
| *Level 1 (Time)* | | | |
| Temporal variation | 0.38 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.23 | *** | 743 |
| Growth rate | 0.03 | ** | 743 |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.20 | *** | 25 |
| Growth rate | 0.03 | *** | 25 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 6.** Variance decompositions for unconditional, adjusted, and final models for general algebra1.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.38 | 0.38 | 0.38 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.26 | 0.23 | 0.23 |
| Growth rate | 0.03 | 0.03 | 0.03 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.35 | 0.22 | 0.20 |
| Growth rate | 0.03 | 0.03 | 0.03 |
| | | | |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.10 | 0.00 |
| Growth rate | -- | 0.06 | 0.01 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.38 | 0.09 |
| Growth rate | -- | 0.04 | 0.17 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.58 | -- | -- |
| Growth rate | 0.53 | -- | -- |

# APPENDIX D

# RESULTS FROM ALL GEOMETRY GROWTH MODELS

## D.1 RESULTS FROM GEOMETRY TIER1 MODEL

**Table 1.** Analysis of tier1 students' geometry achievement with three level hierarchical growth model.

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.04 | | 0.21 |
| *Classroom Level* | | | |
| MSP-PD | 0.18 | | 0.23 |
| Strata1 | 0.17 | | 0.24 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.14 | * | 0.07 |
| Black | -0.31 | * | 0.14 |
| White | 0.12 | | 0.14 |
| Hispanic | -0.18 | | 0.29 |
| Asian | 0.29 | | 0.21 |
| Female | 0.04 | | 0.06 |
| **Linear growth rate** | 0.00 | | 0.11 |
| *Classroom Level* | | | |
| MSP-PD | 0.01 | | 0.11 |
| Strata1 | 0.24 | ~ | 0.11 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | 0.00 | | 0.04 |
| White | 0.00 | | 0.08 |
| Black | 0.00 | | 0.08 |
| Hispanic | 0.03 | | 0.16 |
| Asian | -0.11 | | 0.12 |
| Female | -0.01 | | 0.03 |

| **Random Effects** | *Variance Component* | | *df* |
|---|---|---|---|
| *Level 1 (Time)* | | | |
| Temporal variation | 0.40 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.37 | *** | 795 |
| Growth rate | -- | | -- |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.15 | *** | 11 |
| Growth rate | 0.03 | *** | 11 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 2.** Variance decompositions for unconditional, adjusted, and final models for geometry tier1.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.40 | 0.40 | 0.40 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.36 | 0.31 | 0.31 |
| Growth rate | 0.00 | 0.00 | 0.00 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.23 | 0.16 | 0.15 |
| Growth rate | 0.05 | 0.03 | 0.03 |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.14 | 0.00 |
| Growth rate | -- | 0.00 | 0.08 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.28 | 0.05 |
| Growth rate | -- | 0.30 | 0.00 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.38 | -- | -- |
| Growth rate | 0.95 | -- | -- |

## D.2 RESULTS FROM GEOMETRY TIER2 MODEL

**Table 3.** Analysis of tier2 students' geometry achievement with three level hierarchical growth model.

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | 0.28 | | 0.34 |
| *Classroom Level* | | | |
| MSP-PD | 0.58 | | 0.28 |
| Strata1 | -0.70 | ~ | 0.28 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.09 | | 0.15 |
| Black | -0.15 | | 0.27 |
| White | -0.07 | | 0.28 |
| Hispanic | -0.86 | | 0.65 |
| Asian | -0.48 | | 0.61 |
| Female | 0.25 | * | 0.12 |
| **Linear growth rate** | -0.14 | | 0.19 |
| *Classroom Level* | | | |
| MSP-PD | 0.27 | * | 0.09 |
| Strata1 | -0.13 | | 0.10 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.06 | | 0.10 |
| White | -0.03 | | 0.18 |
| Black | 0.04 | | 0.19 |
| Hispanic | -0.57 | | 0.42 |
| Asian | -0.08 | | 0.42 |
| Female | 0.23 | ** | 0.08 |

| **Random Effects** | *Variance Component* | | *df* |
|---|---|---|---|
| *Level 1 (Time)* | | | |
| Temporal variation | 0.58 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.20 | *** | 232 |
| Growth rate | -- | | -- |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.11 | *** | 4 |
| Growth rate | 0.00 | *** | 4 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 4.** Variance decompositions for unconditional, adjusted, and final models for geometry tier2.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.59 | 0.58 | 0.57 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.14 | 0.08 | 0.14 |
| Growth rate | 0.01 | 0.02 | 0.01 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.31 | 0.20 | 0.11 |
| Growth rate | 0.03 | 0.02 | 0.00 |
| | | | |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.40 | 0.00 |
| Growth rate | -- | 0.00 | 0.71 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.36 | 0.47 |
| Growth rate | -- | 0.11 | 0.84 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.69 | -- | -- |
| Growth rate | 0.84 | -- | -- |

## D.3 RESULTS FROM GENERAL GEOMETRY MODEL

**Table 5.** Analysis of students' geometry achievement across tier1 and tier2 courses with three level hierarchical growth model (Combined model).

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.18 | | 0.25 |
| *Classroom Level* | | | |
| MSP-PD | 0.22 | | 0.17 |
| Strata1 | 0.28 | | 0.26 |
| Strata2 | 0.50 | | 0.28 |
| Strata3 | -0.60 | ~ | 0.29 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.13 | * | 0.06 |
| Black | -0.26 | * | 0.12 |
| White | 0.09 | | 0.12 |
| Hispanic | -0.29 | | 0.26 |
| Asian | 0.20 | | 0.19 |
| Female | 0.09 | ~ | 0.05 |
| **Linear growth rate** | -0.02 | | 0.13 |
| *Classroom Level* | | | |
| MSP-PD | 0.11 | | 0.08 |
| Strata1 | -0.08 | | 0.12 |
| Strata2 | 0.14 | | 0.13 |
| Strata3 | -0.11 | | 0.14 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | 0.00 | | 0.03 |
| White | -0.01 | | 0.07 |

| | | | |
|---|---|---|---|
| Black | 0.01 | | 0.07 |
| Hispanic | -0.05 | | 0.15 |
| Asian | -0.09 | | 0.11 |
| Female | 0.03 | | 0.03 |
| **Random Effects** | *Variance Component* | | *df* |
| *Level 1 (Time)* | | | |
| Temporal variation | 0.39 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.30 | *** | 960 |
| Growth rate | -- | | -- |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.12 | *** | 15 |
| Growth rate | 0.03 | *** | 15 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 6.** Variance decompositions for unconditional, adjusted, and final models for general geometry.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.38 | 0.38 | 0.38 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.27 | 0.24 | 0.24 |
| Growth rate | 0.00 | 0.01 | 0.00 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.35 | 0.13 | 0.12 |
| Growth rate | 0.04 | 0.03 | 0.02 |
| | | | |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.14 | 0.00 |
| Growth rate | -- | 0.00 | 0.18 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.62 | 0.10 |
| Growth rate | -- | 0.27 | 0.13 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.56 | -- | -- |
| Growth rate | 0.91 | -- | -- |

# APPENDIX E

## RESULTS FROM ALL ALGEBRA2 GROWTH MODELS

## E.1 RESULTS FROM ALGEBRA2 TIER1 MODEL

**Table 1.** Analysis of tier1 students' algebra2 achievement with three level hierarchical growth model.

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.23 | * | 0.13 |
| *Classroom Level* | | | |
| MSP-PD | -0.01 | | 0.16 |
| Strata1 | 0.41 | * | 0.16 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.11 | | 0.08 |
| Black | -0.03 | | 0.17 |
| White | 0.38 | * | 0.17 |
| Hispanic | 0.51 | | 0.32 |
| Asian | 0.74 | ** | 0.27 |
| Female | 0.09 | | 0.07 |
| **Linear growth rate** | 0.01 | | 0.04 |
| *Classroom Level* | | | |
| MSP-PD | 0.00 | | 0.06 |
| Strata1 | 0.01 | | 0.05 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.10 | * | 0.04 |
| White | 0.00 | | 0.09 |
| Black | -0.03 | | 0.10 |
| Hispanic | -0.28 | | 0.18 |
| Asian | 0.03 | | 0.15 |
| Female | 0.12 | ** | 0.04 |
| **Random Effects** | *Variance Component* | | *df* |

| | | | |
|---|---|---|---|
| *Level 1 (Time)* | | | |
| Temporal variation | 0.34 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.48 | *** | 641 |
| Growth rate | 0.04 | *** | 641 |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.08 | *** | 12 |
| Growth rate | 0.00 | * | 12 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 2.** Variance decompositions for unconditional, adjusted, and final models for algebra2 tier1.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.34 | 0.34 | 0.34 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.53 | 0.48 | 0.48 |
| Growth rate | 0.04 | 0.04 | 0.04 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.17 | 0.08 | 0.08 |
| Growth rate | 0.01 | 0.00 | 0.00 |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.09 | 0.00 |
| Growth rate | -- | 0.13 | 0.00 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.54 | 0.00 |
| Growth rate | -- | 0.38 | 0.00 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.24 | -- | -- |
| Growth rate | 0.13 | -- | -- |

## E.2 RESULTS FROM ALGEBRA2 TIER2 MODEL

**Table 3.** Analysis of tier2 students' algebra2 achievement with three level hierarchical growth model.

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.09 | | 0.21 |
| *Classroom Level* | | | |
| MSP-PD | -0.17 | | 0.36 |
| Strata1 | 0.58 | | 0.35 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | 0.02 | | 0.13 |
| Black | 0.23 | | 0.28 |
| White | 0.21 | | 0.29 |
| Hispanic | -0.04 | | 0.44 |
| Asian | 0.74 | * | 0.36 |
| Female | 0.10 | | 0.11 |
| **Linear growth rate** | 0.01 | | 0.12 |
| *Classroom Level* | | | |
| MSP-PD | -0.12 | | 0.21 |
| Strata1 | 0.14 | | 0.20 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.02 | | 0.08 |
| White | 0.08 | | 0.17 |
| Black | -0.06 | | 0.18 |
| Hispanic | -0.17 | | 0.28 |
| Asian | -0.05 | | 0.23 |
| Female | 0.05 | | 0.07 |

| **Random Effects** | *Variance Component* | | *df* |
|---|---|---|---|
| *Level 1 (Time)* | | | |
| Temporal variation | 0.75 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.30 | *** | 277 |
| Growth rate | -- | | -- |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.21 | *** | 7 |
| Growth rate | 0.07 | *** | 7 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 4.** Variance decompositions for unconditional, adjusted, and final models for algebra2 tier2.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.55 | 0.55 | 0.55 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.26 | 0.24 | 0.24 |
| Growth rate | 0.02 | 0.02 | 0.02 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.26 | 0.21 | 0.21 |
| Growth rate | 0.07 | 0.07 | 0.07 |
| | | | |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.06 | 0.00 |
| Growth rate | -- | 0.22 | 0.02 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.18 | 0.02 |
| Growth rate | -- | 0.00 | 0.05 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.50 | -- | -- |
| Growth rate | 0.78 | -- | -- |

## E.2 RESULTS FROM GENERAL ALGEBRA2 MODEL

**Table 5.** Analysis of students' algebra2 achievement across tier1 and tier2 courses with three level hierarchical growth model (Combined model).

| **Fixed Effects** | *Coefficient* | | *se* |
|---|---|---|---|
| **Intercept at last time point** | -0.07 | | 0.14 |
| *Classroom Level* | | | |
| MSP-PD | -0.05 | | 0.16 |
| Strata1 | -0.53 | * | 0.21 |
| Strata2 | 0.00 | | 0.24 |
| Strata3 | 0.45 | * | 0.19 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.07 | | 0.06 |
| Black | 0.04 | | 0.14 |
| White | 0.34 | * | 0.14 |
| Hispanic | 0.29 | | 0.25 |
| Asian | 0.72 | ** | 0.20 |
| Female | 0.08 | | 0.06 |
| **Linear growth rate** | 0.04 | | 0.07 |
| *Classroom Level* | | | |
| MSP-PD | -0.04 | | 0.08 |
| Strata1 | -0.13 | | 0.10 |
| Strata2 | -0.01 | | 0.13 |
| Strata3 | 0.01 | | 0.10 |
| *Student Level* | | | |
| Eligible for free/reduced lunch | -0.07 | * | 0.04 |

| | | | |
|---|---|---|---|
| White | 0.02 | | 0.08 |
| Black | -0.03 | | 0.08 |
| Hispanic | -0.22 | | 0.14 |
| Asian | 0.00 | | 0.12 |
| Female | 0.10 | ** | 0.03 |
| **Random Effects** | *Variance Component* | | *df* |
| *Level 1 (Time)* | | | |
| Temporal variation | 0.37 | | -- |
| *Level 2 (Students)* | | | |
| Final status | 0.40 | *** | 888 |
| Growth rate | 0.03 | *** | 888 |
| *Level 3 (Classrooms)* | | | |
| Final status | 0.12 | *** | 20 |
| Growth rate | 0.03 | *** | 20 |
| ~ p-value<.10; * p<.05; ** p<.01; *** p<.001 | | | |

**Table 6.** Variance decompositions for unconditional, adjusted, and final models for general algebra2.

| | Unconditional Model | Adjusted Model (Strata + Student Characteristics) | Final Adjusted Model (MSP-PD) |
|---|---|---|---|
| **Variance Decomposition** | | | |
| Temporal variation (Level 1) | 0.37 | 0.37 | 0.37 |
| *Between students variation (Level 2)* | | | |
| Final Status | 0.43 | 0.40 | 0.40 |
| Growth rate | 0.04 | 0.03 | 0.03 |
| *Between classrooms variation (Level 3)* | | | |
| Final Status | 0.27 | 0.12 | 0.12 |
| Growth rate | 0.03 | 0.03 | 0.03 |
| | | | |
| **Proportion of Variance Explained** | | | |
| *Between students (Level 2)* | | | |
| Final Status | -- | 0.07 | 0.00 |
| Growth rate | -- | 0.12 | 0.00 |
| *Between classrooms (Level 3)* | | | |
| Final Status | -- | 0.57 | 0.01 |
| Growth rate | -- | 0.10 | 0.02 |
| **Intraclass Correlation Coefficient (ICC)** | | | |
| Final Status | 0.39 | -- | -- |
| Growth rate | 0.44 | -- | -- |

## BIBLIOGRAPHY

American Educational Research Association [AERA], (2005). Teaching Teachers: Professional Development To Improve Student Achievement. *Research Points, Essential information for Education Policy, 3*(1), 1-4.

Arbaugh, F., & Brown, C. A. (2005). Analyzing mathematical tasks: A catalyst for change? *Journal of Mathematics Teacher Education, 8*(6), 499–536.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is: or might be: the role of curriculum materials in teacher learning and instructional reform? *Educational researcher*, *25*(9), 6-14.

Ball, D.L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession* (pp. 3-32). San Francisco: Jossey-Bass.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education,60*(5), 497-511.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator, 29*(1), p. 14-17, 20-22, 43-46.

Ball, D. L., Simons, J., Wu, H. H., Simon, R., Whitehurst, G., & Yun, J. (2008). Chapter 5: Report of the task group on teachers and teacher education (pp. 5-i–5-67). Washington, DC: U.S. Department of Education, National Mathematics Advisory Panel.

Berends, M., Kirby, S. N., Naftel, S., & McKelvey, C. (2001). *Implementation and performance in New American Schools: Three Years into scale-up*. Santa Monica, CA: RAND.

Betebenner, D. W., & Linn, R. L. (2010). Growth in student achievement: Issues of measurement, longitudinal data analysis and accountability. Princeton, NJ: K-12 assessment and Performance Management Center, Educational Testing Services.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The elementary school journal*, *111*(1), 7-34.

Birman, B. F., Desimone, L., Garet, M. S., Porter, A. C., & Yoon, K. S. (2001) What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal. 38*(4), 915-945.

Birman, B. F., & Porter, A. C. (2002). Evaluating the effectiveness of education funding streams. *Peabody Journal of Education*, *77*(4), 59-85.

Blank, R. K. & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement: How meta analysis provides scientific evidence useful to education leaders*. Washington, DC: The Council of Chief State School Officers.

Borko, H. (2004). Professional Development and Teacher Learning: Mapping the Terrain. *Educational Researcher, 33*(8), 3-15.

Borko, H., & Putnam, R. (1995). Expanding a teacher's knowledge base: A cognitive psychological perspective on professional development. In T. Guskey and M. Huberman

(Eds.), *Professional development in education* (pp. 35-65). New York: Teachers College Press.

Borman, K., Boydston, T., Lee, R., Lanehart, R., & Cotner, B. (2009, March). *Improving elementary science instruction and student achievement: The impact of a professional development program.* Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects.*Journal of Research on Educational Effectiveness*, *1*(4), 237-264.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147.

Burton, L. (2004). *Mathematicians as Enquirers: Learning about Learning Mathematics*. Dordrecht: Kluwer.

Carpenter, T.P., Fennema, E., Peterson, P., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: an experimental study. *American Educational Research Journal, 26*(4), 499–531.

Chapin, S. (1994). Implementing reform in school mathematics. *Journal of Education, 176*(1), 67–76.

Clewell, B. C., Campbell, P. B., & Perlman, L. (2004). *Review of evaluation studies of mathematics and science curricula and professional development models.* Submitted to the GE Foundation. Washington, DC: Urban Institute.

Coalition for Evidence-Based Policy. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. [Electronic resource].Washington,

DC: U.S. Dept. of Education, Institute of Education Sciences, NationalCenter for Education Evaluation and Regional Assistance.

Cohen, D. K., & Hill, H. C. (2000). Instructional Policy and Classroom Performance: The Mathematics Reform in California. *Teachers College Record*, *102*(2), 294-343.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.

Corcoran, T. B., (1995). Helping teacher teach well: Transforming professional development.*CPRE Policy Briefs, 16*, 69-79.

Correnti, R. (2007). An Empirical Investigation of Professional Development Effects on Literacy Instruction Using Daily Logs. *Educational Evaluation and Policy Analysis, 29* (4), 262-295.

Cuoco, A., Goldenberg, E.P. and Mark, J. (1996). 'Habits of mind: An organizing principle for mathematics curricula', *Journal of Mathematical Behavior, 15*, 375–402.

Cuoco, A. (2001). Mathematics for teaching. *Notices of the American Mathematical Society, 48*, 168-174.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence.* Retrieved from Center for the Study of Teaching and Policy website: http://depts.washington.edu/ctpmail/PDFs/LDH_1999.pdf

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad.* Dallas, TX: National Staff Development Council.

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development:Toward Better Conceptualizations and Measures. *Educational Researcher*, *38*(3), 181 -199.

Epstein, D., & Miller, R. T. (2011). Slow off the mark: Elementary school teachers and the crisis in STEM education. *Education Digest: Essential Readings Condensed for Quick Review, 77*(1), 4-10.

Ferrini-Mundy, J., Burrill, G., Floden, R., & Sandow, D. (2003, April). *Teacher knowledge for teaching school algebra: Challenges in developing an analytical framework*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ferrini-Mundy, J., McCrory, R., & Senk, S. (2006, April). *Knowledge of algebra teaching: Framework, item development, and pilot results*. Paper presented Research symposium at the research presession of NCTM annual meeting. St. Louis, MO

Fishman, B. J., Marx, R. W., Best, S., & Tal, R. T. (2003). Linking teacher and student learning to improve professional development in systemic reform. *Teaching and teacher education*, *19*(6), 643-658.

Floden, R. E., & McCrory, R. (2007, January). *Mathematical knowledge for teaching algebra: Validating an assessment of teacher knowledge*. Paper presented at 11th AMTE annual conference, Irvine, CA.

Garet, M., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . Sztejnberg, L. (2008). *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement* (NCEE 2008-4030). Washington, DC: National Center for Education

Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American educational research journal, 38*(4), 915-945.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., . . . Doolittle, F. (2010). *Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., . . . Doolittle, F. (2011). *Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation* (NCEE 2011-4024). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hamilton, L. S., McCaffrey, D., Klein, S. P., Stecher, B. M., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. *Educational Evaluation and Policy Analysis, 25*(1), 1-29.

Harrell, F. J. (2013). Rms: regression modeling strategies. R package version 3.6-3. 2013.

Harris, K.R., Lane, K., Graham, S., Driscoll, S., Sandmel, K., Brindle, M., & Schatschneider, C. (2012). Practice-based professional development for strategies instruction in writing: A randomized controlled study. *Journal of Teacher Education 63*(2), 103-119.

Harris, D.N. & Sass, T.R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*, 798–812.

Hawley, W., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice (pp 127-150)*. San Francisco:Jossey-Bass.

Heller J. I. (2012). *Effects of Making Sense of SCIENCE™ professional development on the achievement of middle school students, including English language learners* (NCEE 2012-4002). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333-362.

Hill, H. C. (2007). Learning in the teaching workforce. *The Future of Children*, *17*(1), 111-127.

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for research in mathematics education*, 330-351.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American educational research journal, 42*(2), 371-406.

Kelcey, B., & Phelps, G. (2013). Considerations for Designing Group Randomized Trials of Professional Development with Teacher Knowledge Outcomes. *Educational Evaluation and Policy Analysis*, 35.

Kennedy, M. (1998). *Form and substance of inservice teacher education* (Research Monograph No. 13). Madison, WI:National Institute for Science Education, University of Wisconsin–Madison.

Kilpatrick, J., Swafford, J. and Findell, B. (2001) *Adding it up: Helping Children Learn Mathematics*, National Academy Press, Washington, DC.

Kisa, Z., & Correnti, R. (2012). *Examining Implementation Fidelity in America's Choice (AC) Schools: A Longitudinal Analysis of Changes in Professional Development Associated with Changes in Teacher Practice*. Poster presented at 2012 annual meeting of the American Educational Research Association (AERA), Vancouver, BC, Canada.

Knapp, M. S. (2003). Professional development as policy pathway. *Review of Research in Education,* 27(1), 109–157.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology, 4.*

Leinhardt, G. (1980). Modeling and measuring educational treatment in evaluation. *Review of Educational Research*, *50*(3), 393.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *American Psychologist,48*(12), 1181–1209.

Mayer, D. (1999). Measuring Instructional Practice: Can Policymakers Trust Survey Data? *Educational Evaluation and Policy Analysis, 21*(1), 29-45.

Matsumura, L. C., Garnier, H. E., Correnti, R., Junker, B., & Bickel, D. D. (2010). Investigating the effectiveness of a comprehensive literacy coaching program in schools with high teacher mobility. *The Elementary School Journal*, *111*(1), 35-62.

Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model.*Learning and Instruction*, *25*, 35-48.

McMeeking, S., Orsi, R., & Cobb, R. B. (2012). Effects of a teacher professional development program on the mathematics achievement of middle school students. *Journal for Research in Mathematics Education*,*43*(2), 159-181.

McMillan, J. H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment Research & Evaluation*, *12*(15).

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance*, *30*(1), 1–26.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press. Chicago.

National Center for Education Statistics (NCES), (2013). *National Assessment of Educational Progress (NAEP), various years, 1992–2013, Mathematics and Reading Assessments*. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://nationsreportcard.gov/reading_math_g12_2013/.

National Science Foundation (NSF), (2010). *Math and Science Partnership Program: Strengthening America by advancing academic achievement in mathematics and science*. Arlington, VA. Retrieved from http://www.nsf.gov/pubs/2010/nsf10046/nsf10046.pdf

Newman, D., Finney, P.B., Bell, S., Turner, H., Jaciw, A.P., Zacamy, J.L., & Feagans Gould, L. (2012). Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). (NCEE 2012–4008). Washington, DC: National Center for Education

Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston : Pearson Education.

Parsad, B., Lewis, L., & Farris, E. (2001). *Teacher preparation and professional development:2000*(NCES 2001-088). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology*, *102*(2), 299.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206-230.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. 2nd edition. Newbury Park, CA: Sage.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 17*, 41-55.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching*, *48*(2), 117-148.

Rowan, B., Correnti, R., & Miller, R. (2002). What Large-Scale Survey Research Tells Us About Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools. *The Teachers College Record, 104*(8), 1525-1567.

Sailors, M., & Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. *The Elementary School Journal*, *110*(3), 301-322.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Saxe, G.B., Gearhart, M., & Nasir, N.S. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Education, 4*(1), 55–79.

Scher, L., & O'Reilly, F. (2009). Professional Development for K–12 Math and Science Teachers: What Do We Really Know?. *Journal of Research on Educational Effectiveness*, *2*(3), 209-249.

Schwille, A. Porter, G. Belli, R. Floden, D. Freeman, L. Knappen, T. Kuhs, & W. Schmidt (1983). Teachers as policy brokers in the content of elementary school mathematics. In L. Shulman & G. Sykes (eds.), *Handbook of Teaching and Policy*. New York: Longman.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454-499.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education*. National Academies Press.

Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2001). Application of shrinkage techniques in logistic regression analysis: a case study. Statistica Neerlandica, 55(1), 76-88.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., . . . Donald L.C. (2008). Scaling Up an Early Reading Program: Relationships Among Teacher Support, Fidelity of Implementation, and Student Performance Across Different Sites and Years. *Educational Evaluation and Policy Analysis, 30*(4), 368 -388.

Stein, M. K., Smith, M. S., & Silver, E. A. (1999). The development of professional developers: Learning to assist teachers in new settings in new ways. Harvard Educational Review, 69(3), 237–269

Sykes, G., & Darling-Hammond, L., (1999). *Teaching as the learning profession: Handbook of policy and practice*. San Francisco: Jossey-Bass Publishers.

Supovitz, J. A., Mayer, D., & Kahle, J. B. (2000). The longitudinal impact of inquiry-based professional development on teaching practice. *Educational Policy*, *14*(3), 331–356.

Tan, F.E.S. (2008). Best practices in analysis of longitudinal data: a multilevel approach. In J.Osborn (Ed.), *Best practice in quantitative methods* (pp. 451-471). California: Sage.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: motives and methods. Educational Researcher, 37(8), 469-479.